



Docket No.: PF-0505-2 DIV

**Response Under 37 C.F.R. 1.116 - Expedited Procedure**  
**Examining Group 1644**

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to:  
Mail Stop Appeal Brief Patents, Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450 on October 14, 2003.  
By: [Signature] Printed: Katherine Stofer

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**  
**BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES**

In re Application of: Hillman et al.

Title: GROWTH ASSOCIATED PROTEASE INHIBITOR HEAVY CHAIN  
PRECURSOR

Serial No.: 09/828,423

Filing Date: April 05, 2001

Examiner: DeCloux, A.

Group Art Unit: 1644

RECEIVED  
OCT 20 2003  
TECH CENTER 1600/2900

Mail Stop Appeal Brief-Patents  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, Virginia 22313-1450

**BRIEF ON APPEAL**

Sir:

Further to the Notice of Appeal filed June 11, 2003, and received by the USPTO on June 13, 2003, herewith are three copies of Appellants' Brief on Appeal. Authorized fees include the statutory fee of \$420 for a two month extension of time, as well as the \$ 330.00 fee for the filing of this Brief.

This is an appeal from the decision of the Examiner finally rejecting claims 3, 5-6, 8, 11-12, 14-17 and 20-21 of the above-identified application.

**(1) REAL PARTY IN INTEREST**

The above-identified application is assigned of record to Incyte Pharmaceuticals, Inc. (now Incyte Corporation, formerly known as Incyte Genomics, Inc. (Reel 010232, Frame 0779), which is the real party in interest herein.

10/20/2003 MAHME1 00000066 090108 09828423

01 FC:1402 330.00 DA

113597

09/828,423

(2) RELATED APPEALS AND INTERFERENCES

Appellants, their legal representative and the assignee are not aware of any related appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the instant appeal.

(3) STATUS OF THE CLAIMS

Claims rejected: Claims 3, 5-6, 8, 11-12, 14-17 and 20-21  
Claims allowed: (none)  
Claims canceled: Claims 1 and 2  
Claims withdrawn: Claims 4, 7, 9, 10, 13, 18 and 19  
Claims on Appeal: Claims 3, 5-6, 8, 11-12, 14-17 and 20-21 (A copy of the claims on appeal, as amended, can be found in the attached Appendix).

(4) STATUS OF AMENDMENTS AFTER FINAL

There were no amendments made after final.

(5) SUMMARY OF THE INVENTION

Embodiments of Appellants' invention are directed to an isolated antibody which specifically binds to a polypeptide, human growth-associated protease inhibitor heavy chain precursor, abbreviated as "GAPIP". Appellants' invention includes antibodies which specifically bind growth-associated protease inhibitor heavy chain precursors selected from among polypeptides comprising the amino acid sequence of SEQ ID NO:1 (See the Specification, e.g., at page 3, lines 1-10 and the Sequence Listing), polypeptides comprising a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1, said naturally-occurring amino acid sequence encoding a polypeptide having protease inhibitor activity (See the Specification e.g., at page 16, lines 16-19), and an immunogenic fragment of a polypeptide comprising the amino acid sequence of SEQ ID NO:1 (See, e.g., at page 8, lines 1-5 and page 45, lines 7-8).

As described in the Specification at page 14, line 18 to page 15 line 11:

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:1, as shown in Figures 1A, 1B, 1C, 1D, 1E, 1F, 1G, 1H, 1I, and 1J. GAPIP is 942 amino acids in length and has eight potential N-glycosylation sites at residues N97, N127, N231, N421, N508, N776, N795, and N862; twelve potential casein kinase II phosphorylation sites at residues S17, S28, T112, T129, S158, S269, S354, T410, T581, S592, T676, and S754; two potential glycosaminoglycan attachment sites at residues S213 and S391; seventeen potential protein kinase C phosphorylation sites at residues S55, S70, T112, S175, S182, S213, S337, S354, T416, T458, S535, S559, T581, S611, S620, S651, and T880; one potential tyrosine kinase phosphorylation site at residue Y919; a potential signal peptide sequence from M1 to C14; and a vWFA3 domain, which contains the potential metal-binding site glycine-amino acid-serine-amino acid-serine, from N295 to N440. As shown in Figures 2A, 2B, 2C, 2D, 2E, 2F, and 2G, GAPIP has chemical and structural similarity with human pre-inter- $\alpha$ -trypsin inhibitor (GI 33985; SEQ ID NO:3), human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1 (GI 33989; SEQ ID NO:4), and human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3 (GI 288563; SEQ ID NO:5). In particular, GAPIP and human pre-inter- $\alpha$ -trypsin inhibitor share 28% identity, one potential N-glycosylation site, four potential casein kinase II phosphorylation sites, four potential protein kinase C phosphorylation sites, the potential signal peptide sequence, and the vWFA3 potential metal-binding site glycine-amino acid-serine-amino acid-serine. In addition, GAPIP and human pre-inter- $\alpha$ -trypsin inhibitor heavy chains H1 and H3 share 27% and 23 % identity, respectively, one potential N-glycosylation site, four potential casein kinase II phosphorylation sites, five potential protein kinase C phosphorylation sites, the potential signal peptide sequence, and the vWFA3 potential metal-binding site glycine-amino acid-serine-amino acid-serine. As illustrated by Figure 3, GAPIP and human pre-inter- $\alpha$ -trypsin inhibitor heavy chains share a common phylogenetic heritage. A fragment of SEQ ID NO:2 from about nucleotide 982 to about nucleotide 1011 is useful, for example, for designing oligonucleotides or as a hybridization probe. Northern analysis shows the expression of this sequence in various libraries, at least 63% of which are immortalized or cancerous and at least 26% of which involve immune response. Of particular note is the expression of GAPIP in reproductive, gastrointestinal, nervous, and fetal tissues.

The polypeptides and antibodies of the present invention have a variety of utilities. In particular, they can be used in expression profiling, for toxicology testing, for drug discovery, and for the diagnosis, prevention, and treatment of reproductive, developmental, neoplastic, and immunological disorders. (See the Specification at, e.g., at page 14, lines 5-9; page 32, lines 19-30 and page 38, lines 6-22 )

#### (6) ISSUES

1. Whether claims 3, 5-6, 8, 11-12, 14-17 and 20-21 meet the utility requirement of

35 U.S.C. § 101.

2. Whether one of ordinary skill in the art would know how to use the claimed antibodies, e.g., in toxicology testing, drug development, and to the diagnosis of disease, so as to satisfy the enablement requirement of 35 U.S.C. § 112, first paragraph.
3. Whether claims 3, 5-6, 8, 11-12, 14-17 and 20-21 meet the enablement requirement of 35 U.S.C. § 112, first paragraph with respect to whether one of ordinary skill in the art would be able to make and use antibodies which specifically bind polypeptides comprising naturally occurring amino acid sequences that are at least 90% identical to the amino acid sequence of SEQ ID NO:1.
4. Whether claims 3, 5-6, 8, 11-12, 14-17 and 20-21 meet the written description requirement of 35 U.S.C. § 112, first paragraph.

**(7) GROUPING OF THE CLAIMS**

**As to Issue 1**

All of the claims on appeal are grouped together.

**As to Issue 2**

All of the claims on appeal are grouped together.

**As to Issue 3**

Claims 3, 5, 6, 8, 11, 12 and 14-17 should be considered separately from claims 20 and 21.

**As to Issue 4**

Claims 3, 5, 6, 8, 11, 12 and 14-17 should be considered separately from claims 20 and 21.



(8) APPELLANTS' ARGUMENTS

**Issue 1 – Utility rejection Under 35 U.S.C. § 101**

Claims 3, 5-6, 8, 11-12, 14-17 and 20-21 stand rejected under 35 U.S.C. §§ 101 and 112, first paragraph, based on the allegation that the claimed invention lacks patentable utility. The rejection alleges in particular that “the claimed invention is not supported by either a specific and substantial asserted utility, a credible asserted utility or a well-established utility.” (3/22/03 Office Action, at page 2).

The rejection of claims 3, 5-6, 8, 11-12, 14-17 and 20-21 is improper, as the inventions of those claims have a patentable utility as set forth in the instant specification, and/or a utility well-known to one of ordinary skill in the art.

The invention at issue is identified in the patent application as an antibody that specifically binds to growth-associated protease inhibitor heavy chain precursor (GAPIP), which is a polypeptide encoded by a gene that is expressed in reproductive, gastrointestinal, nervous, and fetal tissues (see the Specification, e.g., at page 16, lines 16-22 and page 28, lines 12-14). The novel polypeptide GAPIP to which the claimed antibody specifically binds is demonstrated in the Specification to be a member of the protease inhibitor family (see the Specification at pages 14-15), whose biological functions include regulation of the activity and effect of proteases and control pathogenesis of proteolytic disorders, and in treatment of HIV (see the Specification at page 2, lines 21-23). As such, the claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which require knowledge of how the polypeptide actually functions. As a result of the benefits of these uses, the claimed invention already enjoys significant commercial success.

The fact that the polypeptide to which the claimed antibody specifically binds is a member of the protease inhibitor family alone demonstrates utility beyond the reasonable probability required by law. Each of the members of this class, regardless of their particular functions, are useful. There is no evidence that any member of this class of polypeptides, let alone a substantial number of them, would not have some patentable utility. It follows that there is a more than substantial likelihood that the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds also have patentable utility, regardless of their

actual function. The law has never required a patentee to prove more.

There is, in addition, direct proof of the utility of the claimed invention. The Declaration of Lars Michael Furness (previously submitted with the Office Response of 11/18/02) describes some of the practical uses of the claimed invention in gene and protein expression monitoring applications as they would have been understood at the time of the patent application. The Furness Declaration describes, in particular, how the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds can be used in protein expression analysis techniques such as 2-D PAGE gels and western blots. Using the claimed invention with these techniques, persons of ordinary skill in the art can better assess, for example, the potential toxic affect of a drug candidate (Furness Declaration at ¶ 10).

The Patent Examiner does not dispute that the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds can be used in 2-D PAGE gels and western blots to perform drug toxicity testing. Instead, the Patent Examiner contends that the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds cannot be useful without precise knowledge of its function. But the law never has required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

As demonstrated by the Furness Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds in the absence of any knowledge as to the precise function of the protein. The uses of the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds for gene expression monitoring applications including toxicology testing are in fact independent of its precise function.

## **I. The Applicable Legal Standard**

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); *Fuller v. Berger*, 120 F.2d 274, 275 (7th Cir. 1903) (test for utility is whether invention "is incapable of serving any beneficial end").

*Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. See *Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a "nebulous expression" such as "biological activity" or "biological properties" that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be "substantial." *Brenner*, 383 U.S. at 534. A "substantial" utility is a practical, "real-world" utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a "well-established" utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examination Procedure at § 706.03(a). Only if there is no "well-established" utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed

to possess it. See *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. See *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If, and only if, the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a “substantial likelihood” of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

**II. Uses of the claimed antibodies for the diagnosis of conditions or disorders characterized by expression of GAPIP, for toxicology testing, and for drug discovery are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph**

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are “well-established” uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application’s specification. These uses are explained, in detail, in the Furness Declaration accompanying this response. Objective evidence further corroborates the credibility of the asserted utilities.

**A. The Specification discloses that the SEQ ID NO:1 is GAPIP and therefore there is an asserted utility for the claimed antibodies**

The Examiner alleges that nowhere is it disclosed in the instant specification that SEQ ID NO:1 is GAPIP and, therefore, technically there is no asserted utility for the claimed antibodies (See 6/17/02 Office Action, at page 3). Such, however, is not the case. For example, in the Brief Description of the Figures, it is explicitly stated that “Figures 1A, 1B, 1C, 1D, 1E, 1F, 1G, 1H, 1I, and 1J show the amino acid sequence (SEQ ID NO:1) and nucleic acid sequence (SEQ ID NO:2) of GAPIP.” See the Specification at page 5, lines 2-3. Similar statements identifying

SEQ ID NO:1 as an amino acid sequence of GAPIP can be found throughout the Brief Description of the Figures at page 5, lines 5-11.

**B. The similarity of the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds to another of undisputed utility demonstrates utility**

Because there is a substantial likelihood that the claimed GAPIP is functionally related to human pre-inter- $\alpha$ -trypsin inhibitor, human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1, and pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3, polypeptides of undisputed utility, there is by implication a substantial likelihood that the claimed polypeptide is similarly useful. Appellants need not show any more to demonstrate utility. See *In re Brana*, 51 F.3d at 1567.

It is undisputed, and readily apparent from the patent application, that the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds shares more than 40 % sequence identity over 70 amino acid residues with human pre-inter- $\alpha$ -trypsin inhibitor, human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1, and pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3. For example, over the 70 amino acid residues from G271 to I340 of SEQ ID NO:1, human pre-inter- $\alpha$ -trypsin inhibitor, human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1, and pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3 are 55%, 48% and 54% identical, respectively. This is more than enough homology to demonstrate a reasonable probability that the utility of human pre-inter- $\alpha$ -trypsin inhibitor, human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1, and pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3 can be imputed to the claimed invention. It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner et. al., Proc. Natl. Acad. Sci. 95:6073-78 (1998). Given homology in excess of 40% over many more than 70 amino acid residues, the probability that the claimed polypeptide is related to human pre-inter- $\alpha$ -trypsin inhibitor, human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1, and pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3 is, accordingly, very high.

The Patent Office must accept the Appellants' demonstration that the homology between the claimed invention and human pre-inter- $\alpha$ -trypsin inhibitor, human pre-inter- $\alpha$ -trypsin inhibitor heavy chain H1, and pre-inter- $\alpha$ -trypsin inhibitor heavy chain H3 demonstrates utility

by a reasonable probability unless the Patent Office can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

**C. The uses of the claimed antibody and the SEQ ID NO:1 polypeptide to which the antibody specifically binds for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer “specific benefits” to the public**

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the Furness Declaration. There is no dispute that the claimed invention is in fact a useful tool in two-dimensional polyacrylamide gel electrophoresis (“2-D PAGE”) analysis and western blots used to monitor protein expression and assess drug toxicity.

The instant application, Serial No. 09/828,423, filed on April 5, 2001 (hereinafter “Hillman ‘423 application”), is a divisional of and claims priority to U.S. application Serial No. 09/388,774 filed September 2, 1999, issued May 8, 2001 as U.S. Patent No. 6,228,991, which is a divisional application and claimed priority to U.S. application Serial No. 09/074,579 filed May 7, 1998, issued December 14, 1999 as U.S. Patent No. 6,001,596 (hereinafter “the Hillman ‘579 application”), all having the identical specification., with the exception of corrected typographical errors and reformatting.

In his Declaration, Mr. Furness explains the many reasons why a person skilled in the art who read the Hillman ‘579 application on May 7, 1998 would have understood that application to disclose the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds to be useful for a number of gene and protein expression monitoring applications, *e.g.*, in 2-D PAGE technologies, in connection with the development of drugs and the monitoring of the activity of such drugs. (Furness Declaration at, *e.g.*, ¶¶ 9-13). Much, but not all, of Mr. Furness’ explanation concerns the use of the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds in the creation of protein expression maps using 2-D PAGE.

2-D PAGE technologies were developed during the 1980's. Since the early 1990's, 2-D PAGE has been used to create maps showing the differential expression of proteins in different cell types or in similar cell types in response to drugs and potential toxic agents. Each expression pattern reveals the state of a tissue or cell type in its given environment, *e.g.*, in the presence or absence of a drug. By comparing a map of cells treated with a potential drug candidate to a map of cells not treated with the candidate, for example, the potential toxicity of a drug can be assessed (see Furness Declaration at ¶ 10).

The claimed invention makes 2-D PAGE analysis a more powerful tool for toxicology and drug efficacy testing. A person of ordinary skill in the art can derive more information about the state or states or tissue or cell samples from 2-D PAGE analysis with the claimed invention than without it. As Mr. Furness explains:

In view of the Hillman '579 application, the Wilkins article, and other related pre-May 7, 1998 publications, persons skilled in the art on May 7, 1998 clearly would have understood the Hillman '579 application to disclose the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds to be useful in 2-D PAGE analyses for the development of new drugs and monitoring the activities of drugs for such purposes as evaluating their efficacy and toxicity . . . . (Furness Declaration, ¶ 10)

\* \* \*

Persons skilled in the art would appreciate that a 2-D PAGE map that utilized the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds would be a more useful tool than a 2-D PAGE map that did not utilize this protein sequence in connection with conducting protein expression monitoring studies on proposed (or actual) drugs for treating reproductive, developmental, neoplastic, and immunological disorders for such purposes as evaluating their efficacy and toxicity. (Furness Declaration, ¶ 12)

Mr. Furness' observations are confirmed in the literature published before the filing of the patent application. Wilkins, for example, describes how 2-D gels are used to define proteins present in various tissues and measure their levels of expression, the data from which is in turn used in databases:

For proteome projects, the aim of [computer-aided 2-D PAGE] analysis . . . is to catalogue all spots from the 2-D gel in a qualitative and if possible quantitative manner, so as to define the number of proteins present and their levels of expression. Reference gel images, constructed from one or more gels, for the basis of two-dimensional gel databases. (Wilkins, Tab C, p. 26).

**D. The use of proteins expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now “well-established”**

The technologies made possible by expression profiling using polypeptides are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, as described by Furness in his Declaration.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett, *et al.*, Differential gene expression in drug metabolism and toxicology: practicalities, problems, and potential, *Xenobiotica* 29:655-691 (July 1999) (Reference No. 1):

Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Reference No. 1, page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir, *et al.*, Microarrays and Toxicology: The Advent of Toxicogenomics, *Molecular Carcinogenesis* 24:153-159 (1999) (Reference No. 2); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, *Toxicology Letters* 112-13:467-471 (2000) (Reference No.3).

The more genes – and, accordingly, the polypeptides they encode -- that are available for use in toxicology testing, the more powerful the technique. Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator, Dr. Cynthia Afshari to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Reference No. 4). Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

In fact, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Recent developments provide evidence that the benefits of this information are already beginning to manifest themselves. Examples include the following:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter



gene, and chromosomal mapping location, to identify the key gene associated with Tangier disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.

- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.
- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be reversed.

**E. Objective evidence corroborates the utilities of the claimed invention**

There is in fact no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. "Real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing all expressed genes (along with the polypeptide translations of those genes). (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Appellants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations.

Incyte sells its database containing the GAPIP sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's databases have proven to be valuable in, for example, the identification and development of drug candidates. As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's discovery of the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

### **III. The Patent Examiner's Rejections Are Without Merit**

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by alleging that the disclosed and well-established utilities for the claimed antibody and the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds are not a specific and substantial asserted utility, credible asserted utility or well-established utility (see 6/17/02 Office Action, at page 2). The Examiner is incorrect both as a matter of law and as a matter of fact.

#### **A. The Precise Biological Role Or Function Of An Expressed Polypeptide Is Not Required To Demonstrate Utility**

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological role" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a 2-D gel or western blot to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on

biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d at 1359, but rather whether the invention provides an "identifiable benefit" in presently available form. *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d at 1366. If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the Furness Declaration (at, *e.g.*, ¶¶ 9-13), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called "throwaway" utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged as much (66 F.R. at 1095):

[T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for the SEQ ID NO:1 polypeptide to which the claimed antibody specifically binds, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

#### **B. Membership in a Class of Useful Products Can Be Proof of Utility**

Despite the uncontradicted evidence that the claimed polypeptide is a member of the protease inhibitor family, whose members indisputably are useful, the Examiner refused to impute the utility of the members of the protease inhibitor family to GAPIP. The Patent

Examiner takes the position that unless Appellants can identify which particular biological function within the class of protease inhibitors is possessed by GAPIP, utility cannot be imputed. ( See 6/17/02 Office Action, at pages 3-4) To demonstrate utility by membership in the class of protease inhibitors, the Examiner would require that all protease inhibitors possess a "common" utility.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether the members of the class possess one utility or many. See *Brenner v. Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a "general" class is insufficient to demonstrate utility only if the class contains a substantial number of useless members. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members, e.g., *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).<sup>1</sup>

The Examiner addresses GAPIP as if the general class in which it is included is not the protease inhibitor family, but rather all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these "general classes" may contain a substantial number of useless members, the protease inhibitor family does not. The protease inhibitor family is sufficiently specific to rule out any reasonable possibility that GAPIP would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the protease inhibitor class of proteins has any, let alone a substantial number, of useless members, the Examiner must

---

<sup>1</sup>At a recent Biotechnology Customer Partnership Meeting, PTO Senior Examiner James Martinell described an analytical framework roughly consistent with this analysis. He stated that when an applicant's claimed protein "is a member of a family of proteins that already are known based upon sequence homology," that can be an effective assertion of utility.

conclude that there is a "substantial likelihood" that the GAPIP encoded by the claimed polypeptide is useful.

Even if the Examiner's "common utility" criterion were correct – and it is not – the protease inhibitor family would meet it. It is undisputed that known members of the protease inhibitor family regulate the activity and effect of proteases. A person of ordinary skill in the art need not know any more about how the claimed invention regulates the activity and effect of proteases to use it, and the Examiner presents no evidence to the contrary. Instead, the Examiner makes the conclusory observation that a person of ordinary skill in the art would need to know whether, for example, any given protease inhibitor regulates the activity and effect of proteases. The Examiner then goes on to assume that the only use for GAPIP absent knowledge as to how this member of the protease inhibitor family actually works is further study of GAPIP itself.

Not so. As demonstrated by Appellants, knowledge that GAPIP is a protease inhibitor is more than sufficient to make it useful for the diagnosis and treatment of cancer and immune disorders. Indeed, GAPIP has been shown to be expressed in cancer and immune cells. The Examiner must accept these facts to be true unless the Examiner can provide evidence or sound scientific reasoning to the contrary. But the Examiner has not done so.

**C. The uses of GAPIP in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the invention itself**

There is no authority for the proposition that use as a tool for research is not a substantial utility. Indeed, the Patent Office itself has recognized that just because an invention is used in a research setting does not mean that it lacks utility (Section 2107.01 of the Manual of Patent Examining Procedure, 8<sup>th</sup> Edition, August 2001, under the heading I. Specific and Substantial Requirements, Research Tools):

Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact "useful" in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified substantial utility and inventions whose asserted utility requires further research to identify or reasonably confirm.

The PTO's actual practice has been, at least until the present, consistent with that approach. It has routinely issued patents for inventions whose only use is to facilitate research, such as DNA ligases, acknowledged by the PTO's Training Materials to be useful.

The subset of research uses that are not "substantial" utilities is limited. It consists only of those uses in which the claimed invention is to be an **object** of further study, thus merely inviting further research on the invention itself. This follows from *Brenner*, in which the U.S. Supreme Court held that a process for making a compound does not confer a substantial benefit where the only known use of the compound was to be the object of further research to determine its use. *Id.* at 535. Similarly, in *Kirk*, the Court held that a compound would not confer substantial benefit on the public merely because it might be used to synthesize some other, unknown compound that would confer substantial benefit. *Kirk*, 376 F.2d at 940, 945. ("What Applicants are really saying to those in the art is take these steroids, experiment, and find what use they do have as medicines.") Nowhere do those cases state or imply, however, that a material cannot be patentable if it has some other, additional beneficial use in research.

Such beneficial uses beyond studying the claimed invention itself have been demonstrated, in particular those described in the Furness Declaration. The Furness Declaration demonstrates that the claimed invention is a tool, rather than an object, of research, and it demonstrates exactly how that tool is used. Without the claimed invention, it would be more difficult to generate information regarding the properties of tissues, cells, drug candidates and toxins apart from additional information about the polypeptide itself.

The claimed invention has numerous other uses as a research tool, each of which alone is a "substantial utility." These include uses drug screening (e.g., Specification at page 38).

**D. The Patent Examiner Failed to Demonstrate That a Person of Ordinary Skill in the Art Would Reasonably Doubt the Utility of the Claimed Invention**

The 6/17/02 Office Action has also set forth the novel theory that the central dogma of molecular biology (*i.e.*, DNA directs transcription of messenger RNA which in turn directs translation of protein) somehow does not apply to the discoveries of the present application. That is, the nucleotide sequence of SEQ ID NO:2 (which encodes the polypeptide of SEQ ID NO:1) was determined from a human uterus cDNA library. That cDNA library in turn was made

from messenger RNA isolated from human tissue. See the Specification, for example, at pages 38-39. Thus, the nucleotide sequences of the present invention are expressed sequences. The 6/17/02 Office Action purports that the existence of an expressed mRNA does not insure that the protein encoded by the mRNA will be translated and, hence, the claimed subject matter lacks patentable utility.

Regulation of gene expression occurs at many levels, including transcription, splicing, polyadenylation, mRNA stability, mRNA transport and compartmentalization, translation efficiency, protein modification, and protein turnover. While steady state mRNA levels are not always directly proportional to the amount of protein produced in a cell, mRNA levels are **routinely** used as an indicator of protein expression. Countless scientific publication have been based on data relating to mRNA levels when the polypeptide encoded by the mRNA was unknown or difficult to detect. Moreover, mRNA levels are **usually** a good indicator of protein levels in a cell. The 6/17/02 Office Action, cites an example of inhibition of translation initiation; however, this example represents a comparatively unusual mechanism of gene regulation. According to B. Lewin [(1997) Genes VI Oxford University Press, Inc. New York, NY]:

Transcription of a gene in the active state is controlled at the stage of initiation, that is, by the interaction of RNA polymerase with its promoter. This is now becoming susceptible to study in the *in vitro* systems... ***For most genes, this is a major control point; probably it is the most common level of regulation.*** [page 847, emphasis added].

But having acknowledged that control of gene expression can occur at multiple stages, and that production of RNA cannot inevitably be equated with production of protein, it is clear that ***the overwhelming majority of regulatory events occur at the initiation of transcription. Regulation of tissue-specific gene transcription lies at the heart of eukaryotic differentiation.*** [pages 847-848, emphasis added]

Thus the question is not whether there is the potential for post-transcriptional regulation of SEQ ID NO:1 expression but whether one skilled in the art would have a reasonable expectation that SEQ ID NO:1 expression correlates with the levels of SEQ ID NO:2 mRNA. Applicants need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner v. Manson*, 383 U.S. 519, 532, 148 USPQ 689 (1966). In the case of the instant invention, one skilled in the art would be imprudent in assuming, *a priori*, that protein levels did

not correspond to mRNA levels and that levels of SEQ ID NO:1 were controlled predominantly in a post-transcriptional manner, thereby dismissing the significance of mRNA levels. Inasmuch as the predictive value of mRNA levels applies to the "utility" of Appellants' invention, Appellants request reversal of the rejection.

**IV. By Requiring the Patent Applicant to Assert a Particular or Unique Utility, the Patent Examination Utility Guidelines and Training Materials Applied by the Patent Examiner Misstate the Law**

There is an additional, independent reason to reverse the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website [www.uspto.gov](http://www.uspto.gov), March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: "specific" utilities, which meet the statutory requirements, and "general" utilities, which do not. The Training Materials define a "specific utility" as follows:

A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as "gene probe" or "chromosome marker" would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between "specific" and "general" utilities by assessing whether the asserted utility is sufficiently "particular," *i.e.*, unique (Training Materials at p.52) as compared to the "broad class of invention." (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) ("With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.").)



Such “unique” or “particular” utilities never have been required by the law. To meet the utility requirement, the invention need only be “practically useful,” *Natta*, 480 F.2d 1 at 1397, and confer a “specific benefit” on the public. *Brenner*, 383 U.S. at 534. Thus incredible “throwaway” utilities, such as trying to “patent a transgenic mouse by saying it makes great snake food,” do not meet this standard. Karen Hall, *Genomic Warfare*, *The American Lawyer* 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where “specific utility” is defined (page 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be “definite,” not particular. *Montedison*, 664 F.2d at 375. Appellant is not aware of any court that has rejected an assertion of utility on the grounds that it is not “particular” or “unique” to the specific invention. Where courts have found utility to be too “general,” it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had “useful biological activity” was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a “particular” type of cancer was determined to satisfy the specificity requirement). “Particularity” is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Supra* § III.B. (*Montedison*, 664 F.2d at 374-75).

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of “general” utilities. As a result, the Training Materials paint with too broad a brush.

Rigorously applied, they would render unpatentable whole categories of inventions heretofore considered to be patentable, and that have indisputably benefitted the public, including the claimed invention. *See supra* § III.B. Thus the Training Materials cannot be applied consistently with the law.

**Issue 2 – Utility rejection under 35 U.S.C. § 112, first paragraph**

The rejection set forth in the 6/17/02 Office Action, is based on the assertions discussed above, i.e., that the claimed invention lacks patentable utility. To the extent that the rejection under § 112, first paragraph, is based on the improper allegation of lack of patentable utility under § 101, it fails for the same reasons.

**Issue 3 – Enablement rejection under 35 U.S.C. § 112, first paragraph**

In addition, claims 3, 5-6, 8, 11-12, 14-17 and 20-21 have been rejected as failing to meet the enablement requirement of 35 U.S.C. § 112, first paragraph, because the Specification allegedly does not describe how to make the claimed antibodies. The Examiner does not dispute that the present application describes how to make an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1. ( See 6/17/02 Office Action, at page 6-7). However, the Examiner alleges that the present disclosure does not describe how to make (a) an antibody which specifically binds to a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1; and (b) an antibody which specifically binds to an immunogenic fragment of the amino acid sequence of SEQ ID NO:1. Such, however, is not the case.

At the outset, note that this rejection should not apply to claim 21, which recites an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1. Also, claim 20 should be considered separately from claims 3, 5, 6, 8, 11 and 14-17. That is, claim 20 recites an antibody which specifically binds to an immunogenic fragment having at least 15 contiguous amino acid residues of a polypeptide comprising the amino acid sequence of SEQ ID NO:1. Separate reasons for patentability of claim 20 are discussed below.

The Examiner does not appear to dispute that conventional methods for making

antibodies could be used to make antibodies which specifically bind to a polypeptide comprising a naturally occurring amino acid sequence at least 90% sequence identical to the amino acid sequence of SEQ ID NO:1 or to immunogenic fragments of SEQ ID NO:1. Instead, the Examiner asserts that the present disclosure is deficient because one of skill in the art would not be able to make the variant polypeptides and immunogenic fragments of SEQ ID NO:1 *per se* and, hence, without the variant polypeptides and immunogenic fragments, one would not be able to make antibodies which specifically bind to those variant polypeptides and immunogenic fragments. On the contrary, the Specification is sufficient in this regard.

Note that claim 3 recites not only that the variant polypeptides are at least 90% identical to SEQ ID NO:1, but also have “*a naturally-occurring amino acid sequence.*” Through the process of natural selection, nature will have determined the appropriate amino acid sequences. Given the information provided by SEQ ID NO:1 (the amino acid sequence of GAPIP) and SEQ ID NO:2 (the polynucleotide sequence encoding GAPIP), one of skill in the art would be able to routinely obtain “a naturally-occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1.” For example, the identification of relevant polynucleotides could be performed by hybridization and/or PCR techniques that were well-known to those skilled in the art at the time the subject application was filed and/or described throughout the Specification of the instant application. For example:

As used herein, the term “stringent conditions” refers to conditions which permit hybridization between polynucleotides and the claimed polynucleotides. Stringent conditions can be defined by salt concentration, the concentration of organic solvent (e.g., formamide), temperature, and other conditions well known in the art. In particular, stringency can be increased by reducing the concentration of salt, increasing the concentration of formamide, or raising the hybridization temperature.

For example, stringent salt concentration will ordinarily be less than about 750 mM NaCl and 75 mM trisodium citrate, preferably less than about 500 mM NaCl and 50 mM trisodium citrate, and most preferably less than about 250 mM NaCl and 25 mM trisodium citrate. Low stringency hybridization can be obtained in the absence of organic solvent, e.g., formamide, while high stringency hybridization can be obtained in the presence of at least about 35% formamide, and most preferably at least about 50% formamide. Stringent temperature conditions will ordinarily include temperatures of at least about 30°C, more preferably of at least about 37°C, and most preferably of at least about 42°C. Varying additional parameters, such as hybridization time, the concentration of detergent, e.g.,

sodium dodecyl sulfate (SDS), and the inclusion or exclusion of carrier DNA, are well known to those skilled in the art. Various levels of stringency are accomplished by combining these various conditions as needed. In a preferred embodiment, hybridization will occur at 30°C in 750 mM NaCl, 75 mM trisodium citrate, and 1% SDS. In a more preferred embodiment, hybridization will occur at 37°C in 500 mM NaCl, 50 mM trisodium citrate, 1% SDS, 35% formamide, and 100 µg/ml denatured salmon sperm DNA (ssDNA). In a most preferred embodiment, hybridization will occur at 42°C in 250 mM NaCl, 25 mM trisodium citrate, 1% SDS, 50 % formamide, and 200 µg/ml ssDNA. Useful variations on these conditions will be readily apparent to those skilled in the art.

The washing steps which follow hybridization can also vary in stringency. Wash stringency conditions can be defined by salt concentration and by temperature. As above, wash stringency can be increased by decreasing salt concentration or by increasing temperature. For example, stringent salt concentration for the wash steps will preferably be less than about 30 mM NaCl and 3 mM trisodium citrate, and most preferably less than about 15 mM NaCl and 1.5 mM trisodium citrate. Stringent temperature conditions for the wash steps will ordinarily include temperature of at least about 25°C, more preferably of at least about 42°C, and most preferably of at least about 68°C. In a preferred embodiment, wash steps will occur at 25°C in 30 mM NaCl, 3 mM trisodium citrate, and 0.1% SDS. In a more preferred embodiment, wash steps will occur at 42°C in 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. In a most preferred embodiment, wash steps will occur at 68°C in 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. Additional variations on these conditions will be readily apparent to those skilled in the art.

(Specification at page 12, line 12 to page 13, line 12)

In one aspect, hybridization with PCR probes which are capable of detecting polynucleotide sequences, including genomic sequences, encoding GAPIP or closely related molecules may be used to identify nucleic acid sequences which encode GAPIP. The specificity of the probe, whether it is made from a highly specific region, e.g., the 5' regulatory region, or from a less specific region, e.g., a conserved motif, and the stringency of the hybridization or amplification (maximal, high, intermediate, or low), will determine whether the probe identifies only naturally occurring sequences encoding GAPIP, allelic variants, or related sequences. (Specification at page 34, lines 2-8)

Probes may also be used for the detection of related sequences, and should preferably have at least 50% sequence identity to any of the GAPIP encoding sequences. The hybridization probes of the subject invention may be DNA or RNA and may be derived from the sequence of SEQ ID NO:2 or from genomic sequences including promoters, enhancers, and introns of the GAPIP gene. (Specification at page 34, lines 9-12)

See also Example VI at pages 44-45.

Thus, one skilled in the art need not make and test vast numbers of polypeptides that are based on the amino acid sequence of SEQ ID NO:1. Instead, one skilled in the art need only screen a cDNA library or use appropriate PCR conditions to identify relevant polynucleotides/polypeptides that already exist in nature. By adjusting the nature of the probe or nucleic acid (*i.e.*, non-conserved, conserved or highly conserved) and the conditions of hybridization (maximum, high, intermediate or low stringency), one can obtain variant polynucleotides of SEQ ID NO:2 which, in turn, will allow one to make the variant polypeptides of SEQ ID NO:1 recited by the present claims. Conventional methods for making antibodies, such as those described at pages 26-28 of the Specification, could then be used to make antibodies which specifically bind to the recited polypeptide variants.

Accordingly, the Abaza et al. document cited by the Examiner relating to structure-function relationships in proteins is simply not germane to whether one can make and use the polypeptide variants recited by the present claims. That is, regardless of the precise functional characteristics of the SEQ ID NO:1 variants, one can still make those polypeptide variants, and antibodies which specifically bind to the variants, using the disclosure provided by the present Specification. The antibodies could then be used in, for example, diagnostic testing, drug discovery, expression profiling, etc. (See, e.g., Furness Declaration).

Furthermore, the Board's attention is also directed to the enclosed reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078) (Reference No. 5). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that  $\geq 40\%$  identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

Claim 3 recites, *inter alia*, antibodies which specifically bind to "a polypeptide comprising . . . a naturally occurring amino acid sequence at least 90% identical to the amino

acid sequence of SEQ ID NO:1.” In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as growth-associated protease inhibitor heavy chain precursors and which have as little as 30% identity over at least 150 residues to SEQ ID NO:1. The “90% variants” recited by the present claims have a variation that is far less than that of all potential growth-associated protease inhibitor heavy chain precursors related to SEQ ID NO:1, i.e., those growth-associated protease inhibitor heavy chain precursors having as little as 30% identity over at least 150 residues to SEQ ID NO:1. Therefore, one would expect the SEQ ID NO:1 variants recited by the present claims to have the functional activities of a growth-associated protease inhibitor heavy chain precursor.

Furthermore, the Examiner has asserted that one of skill in the art could not make and use an isolated antibody which specifically binds an immunogenic fragment of SEQ ID NO:1. Such, however, is not the case.

At pages 14-15, the Specification describes the polynucleotide of SEQ ID NO:2, the polypeptide encoded by that polynucleotide, *i.e.*, SEQ ID NO:1, and chemical and structural characteristics thereof. The polypeptide and fragments thereof can be produced by either recombinant means (see, *e.g.*, the Specification at pages 18-23) or by chemical synthesis (see, *e.g.*, the Specification at page 18, lines 15-26; and page 23, lines 27-31). As discussed at length above in connection with the “utility” rejection, the use of antibodies and the polypeptides to which they specifically bind for diagnosis of diseases, for toxicology testing, and for drug discovery are well known in the art, *e.g.*, via the use of expression profiling. Such uses are also described in the Specification, *e.g.*, at pages 33-38. Hence, the requirement for providing objective enablement has been met.

The Examiner questions in particular, whether the present Specification provides sufficient guidance to enable the identification of immunogenic fragments of SEQ ID NO:1 (See the 6/17/02 Office Action at page 7 ). The Examiner’s concerns are untenable as the Specification is fully sufficient in this regard.

First note that at page 8, lines 5-6, “immunologically active” is defined as “the capability of the natural, recombinant, or synthetic GAPIP, or of any oligopeptide thereof, to induce a specific immune response in appropriate animals or cells and to bind with specific antibodies.” Specific binding is further defined at page 12 as meaning:

... that interaction between a protein or peptide and an agonist, an antibody, or an antagonist. The interaction is dependent upon the presence of a particular structure of the protein, e.g., the antigenic determinant or epitope, recognized by the binding molecule. For example, if an antibody is specific for epitope "A," the presence of a polypeptide containing the epitope A, or the presence of free unlabeled A, in a reaction containing free labeled A and the antibody will reduce the amount of labeled A that binds to the antibody.

Methods of producing specifically binding antibodies are described in the Specification, for example, at pages 26-28. In this regard, note the paragraph at page 27, lines 3-9, which describes fragment sizes of GAPIP for raising antibodies. See also page 48 which describes the production of antibodies to fragments of GAPIP, including the description of how to identify appropriate immunogenic sites of GAPIP:

... the GAPIP amino acid sequence is analyzed using LASERGENE software to determine regions of high immunogenicity, and a corresponding oligopeptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art. (See, e.g., Ausubel supra, ch. 11.) (Specification at page 48, lines 24-28)

Not only is the Examiner's position factually incorrect (as shown above) but it is also legally in error. As set forth in *In re Marzocchi*, 169 USPQ 367, 369 (CCPA 1971):

The first paragraph of § 112 ***requires nothing more than objective enablement.*** [emphasis added] How such a teaching is set forth, either by the use of illustrative examples or by broad terminology, is of no importance.

As a matter of Patent Office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.

Contrary to the standard set forth in *Marzocchi*, the Examiner has failed to provide any ***reasons*** why one would doubt that the guidance provided by the present Specification would enable one to make and use the recited antibodies which specifically bind to the variants and fragments of SEQ ID NO:1. Therefore, a *prima facie* case for non-enablement has not been established.

For at least the above reasons, reversal of this rejection is requested.

**Issue 4 – Written description rejection**

Claims 3, 5-6, 8, 11-12, 14-17 and 20-21 were rejected under 35 U.S.C. § 112, first paragraph, as allegedly being based on a Specification which provides an inadequate written description of what is claimed. The rejection of claims 3, 5-6, 8, 11-12, 14-17 and 20-21 is in error; the claims meet the written description requirement of 35 U.S.C. 112, first paragraph.

The Examiner appears to be taking the position that every single member of the claimed genus of polypeptides “and the antibodies that bind these fragments and variants” must be specifically disclosed by the Specification, otherwise an inadequate written description has been set forth. (See 6/17/02 Office Action at page 5). However this position is erroneous; no such disclosure is required for an adequate written description.

At the outset, note that this rejection should not apply to claim 21, which recites an antibody which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1. Also, claim 20 should be considered separately from claims 3, 5, 6, 8, 11 and 14-17. That is, claim 20 recites an antibody which specifically binds to an immunogenic fragment having at least 15 contiguous amino acid residues of a polypeptide comprising the amino acid sequence of SEQ ID NO:1. Separate reasons for patentability of claim 20 are discussed below.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. § 112, first paragraph, are well established by case law.

... the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession of *the invention*. The invention is, for purposes of the "written description" inquiry, *whatever is now claimed*. *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

The Board's attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that:

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics<sup>42</sup> which provide evidence that applicant was in possession of the claimed invention,<sup>43</sup> i.e., complete or partial structure, other physical and/or chemical properties, functional



characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics.<sup>44</sup> What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail.<sup>45</sup> If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met.<sup>46</sup>

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art

**a. The specification provides an adequate written description of the claimed "variants" and "fragments" of SEQ ID NO:1**

The subject matter encompassed by claims 3, 5-6, 8, 11-12, 14-17 and 20-21 is either disclosed by the Specification or is conventional or well known to one skilled in the art.

First note that the "variant" and "fragment" language of independent claim 23 recites an isolated antibody which specifically binds to a polypeptide comprising "a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1", or "an immunogenic fragment of a polypeptide comprising the amino acid sequence of SEQ ID NO:1." The polypeptide sequence of SEQ ID NO:1 is explicitly disclosed in the specification. See, for example, the Sequence Listing. Variants of SEQ ID NO:1 are described in the Specification at, for example, page 3, lines 4-5; page 6, lines 2-5 and 9-15; and page 15, lines 12-15. Fragments of SEQ ID NO:1 are described in the Specification at, for example, page 3, lines 4-5; page 4, lines 2-4; and page 7, lines 1-7.

One of ordinary skill in the art would recognize polypeptide sequences which are variants at least 90% identical to SEQ ID NO:1. Given any naturally occurring polypeptide sequence, it would be routine for one of skill in the art recognize whether it was a variant of SEQ ID NO:1. Similarly, SEQ ID NO:1 provides the blueprint to describe any immunogenic fragment thereof. Accordingly, the Specification provides an adequate written description of the recited variants and fragments of SEQ ID NO:1.

**1. The present claims specifically define the claimed genus through the recitation of chemical structure**

Court cases in which “DNA claims” have been at issue (which are hence relevant to claims to proteins encoded by the DNA, and antibodies which specifically bind to those proteins) commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

In claims to genetic material, however, a generic statement such as “vertebrate insulin cDNA” or “mammalian insulin cDNA,” without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a

description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. § 112; *i.e.*, “an mRNA of a vertebrate, which mRNA encodes insulin” in *Lilly*, and “DNA which codes for a human fibroblast interferon-beta polypeptide” in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polypeptides in terms of chemical structure, rather than functional characteristics. For example, the language of independent claim 3 recites chemical structure to define the claimed genus:

3. An isolated antibody which specifically binds to a polypeptide selected from the group consisting of:

- a) a polypeptide comprising the amino acid sequence of SEQ ID NO:1,
- b) a polypeptide comprising a naturally-occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1, said naturally-occurring amino acid sequence encoding a polypeptide having protease inhibitor activity, and
- c) an immunogenic fragment of a polypeptide comprising the amino acid sequence of SEQ ID NO:1.

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the claimed antibodies and the polypeptides to which they specifically bind. Moreover, the recitation of functional characteristics (*i.e.*, “having protease inhibitor activity” with respect to the recited variants of SEQ ID NO:1, and “immunogenic” with respect to the recited fragments of SEQ ID NO:1) adds to the structural recitations of the claims. The antibodies and the polypeptides to which they specifically bind that are defined by the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry “on whatever is now claimed,” the Examiner failed to provide an appropriate

analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*.

**2. The present claims do not define a genus which is "highly variant"**

Furthermore, the claims at issue do not describe a genus which could be characterized as "highly variant". Available evidence illustrates that, rather than being a large variable genus, the claimed genus is of narrow scope.

In support of this assertion, the Board's attention is directed to the reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078)(of record). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues (Brenner et al., pages 6073 and 6076). Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that  $\geq 40\%$  identity over at least 70 residues is reliable in signifying homology between proteins (Brenner et al., page 6076).

The present application is directed, *inter alia*, to antibodies which specifically bind to polypeptides related to human growth-associated protease inhibitor heavy chain precursor (GAPIP). In particular, the polypeptides are selected from amino acid sequences comprising SEQ ID NO:1, naturally occurring amino acid sequences at least 90% identical to SEQ ID NO:1, or immunogenic fragments of SEQ ID NO:1. In accordance with Brenner et al., naturally occurring molecules may exist which could be characterized as human growth-associated protease inhibitor heavy chain precursor (GAPIP) proteins and which have as little as 30% identity over at least 150 residues to SEQ ID NO:1. The "variant language" of the present claims recites a polypeptide comprising "a naturally occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1" (note that SEQ ID NO:1 has 942 amino acid residues). This variation is far less than that of all potential GAPIP proteins related to SEQ ID NO:1, i.e., those GAPIP proteins having as little as 30% identity over at least 150 residues to SEQ ID NO:1.

**3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications**

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. § 112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those cases was based on the state of the art at essentially the "dark ages" of recombinant DNA technology.

The present application has a priority date of May 7, 1998. Much has happened in the development of recombinant DNA technology in the 19 or so years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances, one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polypeptide variants and fragments at the time of filing of this application.

**4. Summary**

The Final Office Action of 3/11/03 failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polypeptides defined by the present claims is adequately described, as evidenced by Brenner et al. Furthermore, there have

been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

For at least the reasons set forth above, the Specification provides an adequate written description of the claimed subject matter, and this rejection should be reversed.

(9) CONCLUSION

Appellants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of “lack of specificity” and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Further not only are they scientifically without merit, but are not supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, “like a nose of wax,”<sup>2</sup> to target rejections of claims to polypeptide and polynucleotide sequences, as well as to claims to methods of detecting said polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be reversed.

Moreover, to the extent the above rejections were based on the Revised Interim and final Examination Guidelines and Training Materials, those portions of the Guidelines and Training Materials that form the basis for the rejections should be determined to be inconsistent with the law.

Finally, the enablement and written description rejections based on alleged deficiencies pertaining to antibodies which specifically bind the recited variants and fragments of SEQ ID NO:1 should also be reversed for at least the reasons set forth herein.

---

<sup>2</sup>“The concept of patentable subject matter under §101 is not ‘like a nose of wax which may be turned and twisted in any direction \* \* \*.’ *White v. Dunbar*, 119 U.S. 47, 51.” (*Parker v. Flook*, 198 USPQ 193 (US SupCt 1978))

Due to the urgency of this matter and its economic and public health implications, an expedited review of this appeal is earnestly solicited.

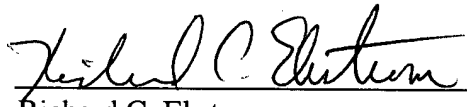
If the USPTO determines that any additional fees are due, the Commissioner is hereby authorized to charge Deposit Account No. 09-0108.

This brief is enclosed in triplicate.

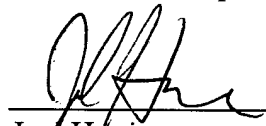
Respectfully submitted,

INCYTE CORPORATION

Date: 14 October 2003

  
Richard C. Ekstrom  
Reg. No. 37,027  
Direct Dial Telephone: (650) 843-7352

Date: Oct 14, 2003

  
Joel Harris  
Reg. No. 44,743  
Direct Dial Telephone: (650) 845-4866

3160 Porter Drive  
Palo Alto, California 94304  
Phone: (650) 855-0555  
Fax: (650) 849-8886

Enclosures:

1. Brenner et al., Proc. Natl. Acad. Sci. U.S.A. 95:6073-78 (1998).
2. John C. Rockett, et. al., Differential gene expression in drug metabolism and toxicology: practicalities, problems, and potential, Xenobiotica 29:655-691 (July 1999).
3. Emile F. Nuwaysir, et al., Microarrays and Toxicology: The Advent of Toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999).
4. Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000).
5. Email from the primary investigator, Dr. Cynthia Afshari to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.

**APPENDIX - CLAIMS ON APPEAL**

3. An isolated antibody which specifically binds to a polypeptide selected from the group consisting of:
- a) a polypeptide comprising the amino acid sequence of SEQ ID NO:1,
  - b) a polypeptide comprising a naturally-occurring amino acid sequence at least 90% identical to the amino acid sequence of SEQ ID NO:1, said naturally-occurring amino acid sequence encoding a polypeptide having protease inhibitor activity, and
  - c) an immunogenic fragment of a polypeptide comprising the amino acid sequence of SEQ ID NO:1.
5. The antibody of claim 3, wherein the antibody is:
- (a) a chimeric antibody;
  - (b) a single chain antibody;
  - (c) a Fab fragment;
  - (d) a F(ab')<sub>2</sub> fragment; or
  - (e) a humanized antibody.
6. A composition comprising an antibody of claim 3 and an acceptable excipient.
8. A composition of claim 6, wherein the antibody is labeled.
11. A polyclonal antibody produced by a method of claim 10.
12. A composition comprising the polyclonal antibody of claim 11 and a suitable carrier.
14. A monoclonal antibody produced by a method of claim 13.
15. A composition comprising the monoclonal antibody of claim 14 and a suitable carrier.



16. The antibody of claim 3, wherein the antibody is produced by screening a Fab expression library.

17. The antibody of claim 3, wherein the antibody is produced by screening a recombinant immunoglobulin library.

20. An isolated antibody of claim 3, which specifically binds to an immunogenic fragment having at least 15 contiguous amino acid residues of a polypeptide comprising the amino acid sequence of SEQ ID NO:1.

21. An isolated antibody of claim 3, which specifically binds to a polypeptide comprising the amino acid sequence of SEQ ID NO:1.

## Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER<sup>\*†‡</sup>, CYRUS CHOTHIA<sup>\*</sup>, AND TIM J. P. HUBBARD<sup>§</sup>

<sup>\*</sup>MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and <sup>†</sup>Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

**ABSTRACT** Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA  $k_{\text{tup}} = 1$ , and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

**Previous Assessments of Sequence Comparison.** Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed ( $k_{\text{tup}} = 2$ ) or greater effectiveness ( $k_{\text{tup}} = 1$ ). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0  
PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPQ, errors per query.

<sup>†</sup>Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

<sup>‡</sup>To whom reprints requests should be addressed. e-mail: [brenner@hyper.stanford.edu](mailto:brenner@hyper.stanford.edu).

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSP equation: it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

**A Database for Testing Homology Detection.** Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ~0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

**Assessment Data and Procedure.** Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0i76 (3), which provided FASTA and the SSEARCH implementation of Smith-Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties -12/-1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

**The "Coverage Vs. Error" Plot.** To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

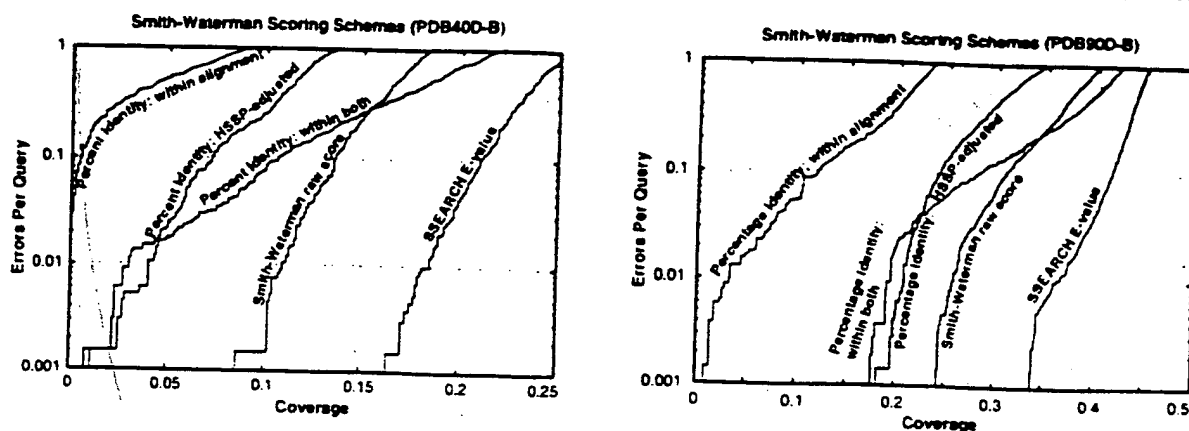


FIG. 1. Coverage vs. error plots of different scoring schemes for sSEARCH Smith-Waterman. (A) Analysis of PDB400-B database. (B) Analysis of PDB400-B database. All of the proteins in the database were compared with each other using the sSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB400-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB400-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSP equation (17) is  $H = 290.15I^{-0.562}$  where  $I$  is length for  $10 < I < 80$ ;  $H > 100$  for  $I < 10$ ;  $H = 24.7$  for  $I > 80$ . The percentage identity HSP-adjusted score is the percent identity within the alignment minus  $H$ . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

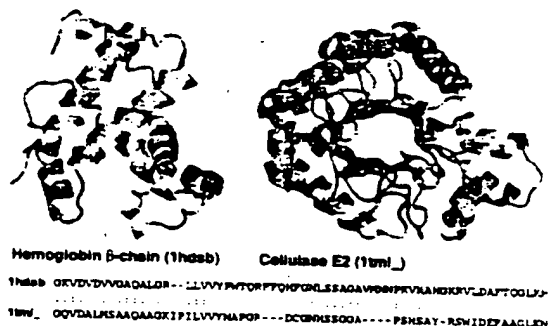


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin  $\beta$ -chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1lm1, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

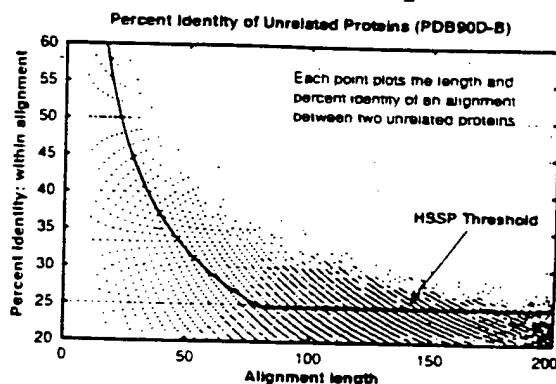


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90-D. Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSP threshold (though it is intended to be applied with a different matrix and parameters).

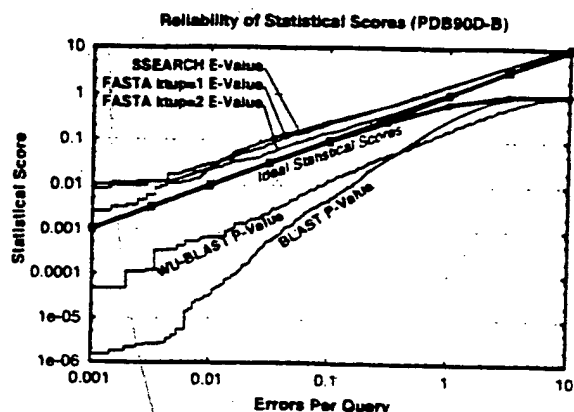


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPO for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPO but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPO measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

**The Performance of Scoring Schemes.** All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

**Sequence Identity.** Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPO. In short, percentage identity discards most of the information measured in a sequence comparison.

**Raw Scores.** Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPO. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

**Statistical Scores.** Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

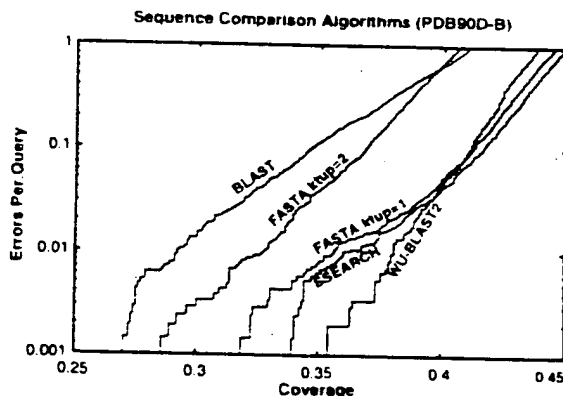
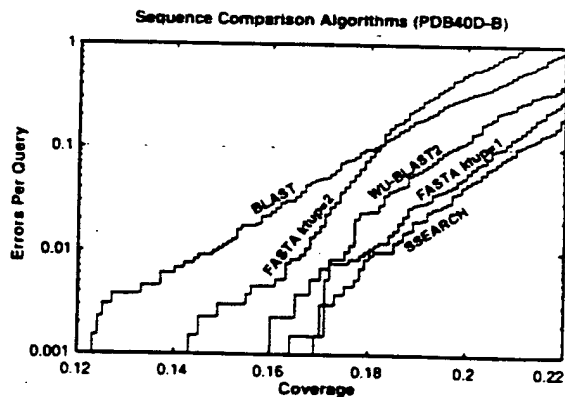


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPO. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPO on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPO for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPO.

**Overall Detection of Homologs and Comparison of Algorithms.** The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPO. BLAST, which identifies 15%, was the worst performer, whereas FASTA  $k_{\text{up}} = 1$  is nearly as effective as SSEARCH. FASTA  $k_{\text{up}} = 2$  and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA  $k_{\text{up}} = 1$ . WU-BLAST2 is slightly faster than FASTA  $k_{\text{up}} = 2$ , but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA  $k_{\text{up}} = 1$ , SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

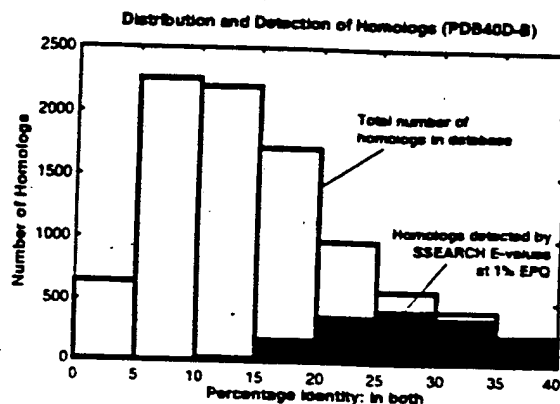


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPO. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

## CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPO Cutoff	Coverage at 1% EPO
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP = 9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA $k_{\text{up}} = 1$ E-values	3.9	0.03	17.9
FASTA $k_{\text{up}} = 2$ E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

\*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.\*\*

\*\*Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403-410.
- Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460-480.
- Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444-2448.
- Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995) *J. Mol. Biol.* 247, 536-540.
- Brenner, S. E., Chothia, C., Hubbard, T. J. P., & Murzin, A. G. (1996) *Methods Enzymol.* 266, 635-643.
- Pearson, W. R. (1991) *Genomics* 11, 635-650.
- Pearson, W. R. (1995) *Protein Sci.* 4, 1145-1160.
- Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195-197.
- George, D. G., Hunt, L. T., & Barker, W. C. (1996) *Methods Enzymol.* 266, 41-59.
- Vogt, G., Etzold, T., & Argos, P. (1995) *J. Mol. Biol.* 249, 816-831.
- Henikoff, S. & Henikoff, J. G. (1993) *Proteins* 17, 49-61.
- Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* 24, 21-25.
- Bairoch, A., Bucher, P., & Hofmann, K. (1996) *Nucleic Acids Res.* 24, 189-196.
- Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* 89, 10915-10919.
- Dayhoff, M., Schwartz, R. M., & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biochemical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345-352.
- Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
- Sander, C. & Schneider, R. (1991) *Proteins* 9, 56-68.
- Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* 233, 716-738.
- Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* 1, 89-94.
- Lesk, A. M., Levitt, M., & Chothia, C. (1986) *Protein Eng.* 1, 77-78.
- Arratia, R., Gordon, L., & M. W. (1986) *Ann. Stat.* 14, 971-993.
- Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* 87, 2264-2268.
- Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* 90, 5873-5877.
- Altschul, S. F., Boguski, M. S., Gish, W., & Wootton, J. C. (1994) *Nat. Genet.* 6, 119-129.
- Pearson, W. R. (1996) *Methods Enzymol.* 266, 227-258.
- Lipman, D. J., Wilbur, W. J., Smith, T. F., & Waterman, M. S. (1984) *Nucleic Acids Res.* 12, 215-226.
- Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* 266, 554-571.
- Waterman, M. S. & Vingron, M. (1994) *Stat. Science* 9, 367-381.
- Perutz, M. F., Kendrew, J. C., & Watson, H. C. (1965) *J. Mol. Biol.* 13, 669-678.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G., & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107-132.
- Brenner, S. E., Chothia, C., & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* 7, 369-376.
- Orengo, C., Michie, A., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. (1997) *Structure (London)* 5, 1093-1108.
- Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* 39, 561-577.
- Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* 20, 25-33.
- Fitch, W. M. (1966) *J. Mol. Biol.* 16, 9-16.
- Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* 4, 1123-1127.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997) *Nucleic Acids Res.* 25, 3389-3402.
- Girling, R., Schmidt, W., Jr., Houston, T., Amma, E., & Huisman, T. (1979) *J. Mol. Biol.* 131, 417-433.
- Spezio, M., Wilson, D., & Karplus, P. (1993) *Biochemistry* 32, 9906-9916.
- Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* 20, 374-376.

## Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡  
and G. GORDON GIBSON\*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,  
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

### Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wyntford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

\* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.



cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrat

n of specific and  
ancer or previous

ymes (including  
le by drugs and  
g transcriptional  
iditi nal cellular  
Accordingly, the  
lement of genes  
development of  
zyme induction  
hemical-induced  
ers reactions to  
me of which are  
phenomenon *per*  
ing technologies  
in target tissues  
is of xenobiotic-  
target tissues is  
enerated in the  
ification of toxic  
nd contributing  
the gene profile  
*in vivo* could be  
of all new drug  
otoxicity, thereby  
such toxicants.  
ality of all genes  
erm goal, as the  
ir functionality  
*pattern* of gene  
d to that of well-  
*in vivo* similarities  
tform for more  
inning to gain  
ially producing  
y assessment of  
me of which are  
y-relat d to any  
broad-spectrum  
arrays are now  
n gr with factor  
nically-induced

ges pres nts a  
velopm nt and  
ing this difficult-  
ods hav been  
at dem nstrate

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

- (1) Differential screening,
- (2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
- (3) Differential display (DD),
- (4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
- (5) Gene expression arrays, and
- (6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences). EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

### Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example; St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

### Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

### Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Duguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

A sequence from a normal, filter replicas are otherwise) labelled cDNA populations. All population will the treated cells. be used to probe which are only up- (1979) screened in order to obtain and breaking in its ming, as up to 2 are differentially ent way to check completed.

Success of early ion gave rise to a be developed was (low). In general, population (tester) eparation of the rihybridized common through the use

Physical separation Several methods chromatography (Dinauer 1990) roach, common om control cells) phy, as hydroxy-sorbed cDNA is ntially expressed ly as a probe to 34, Hedrick *et al.*

Activity enhancing of the problems uid and Dinauer stems as a means the control and or ('oligovector',

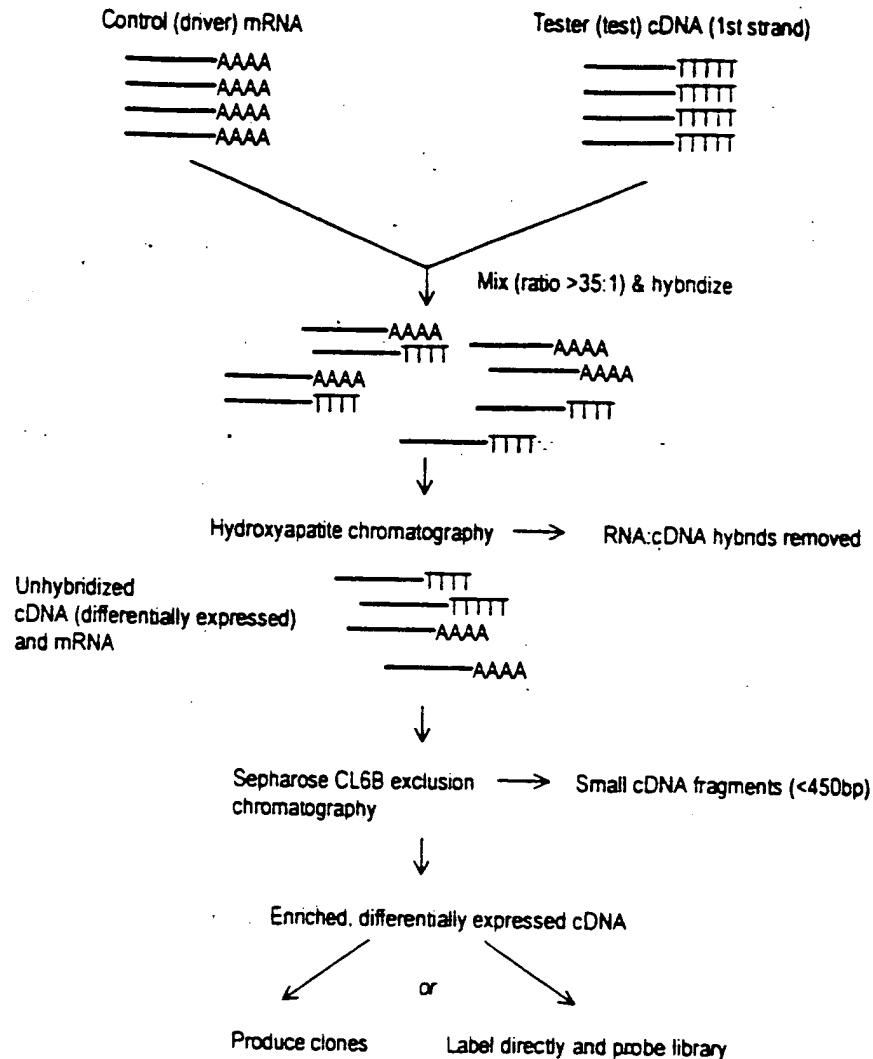


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/alterd (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/alterd population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digest of control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

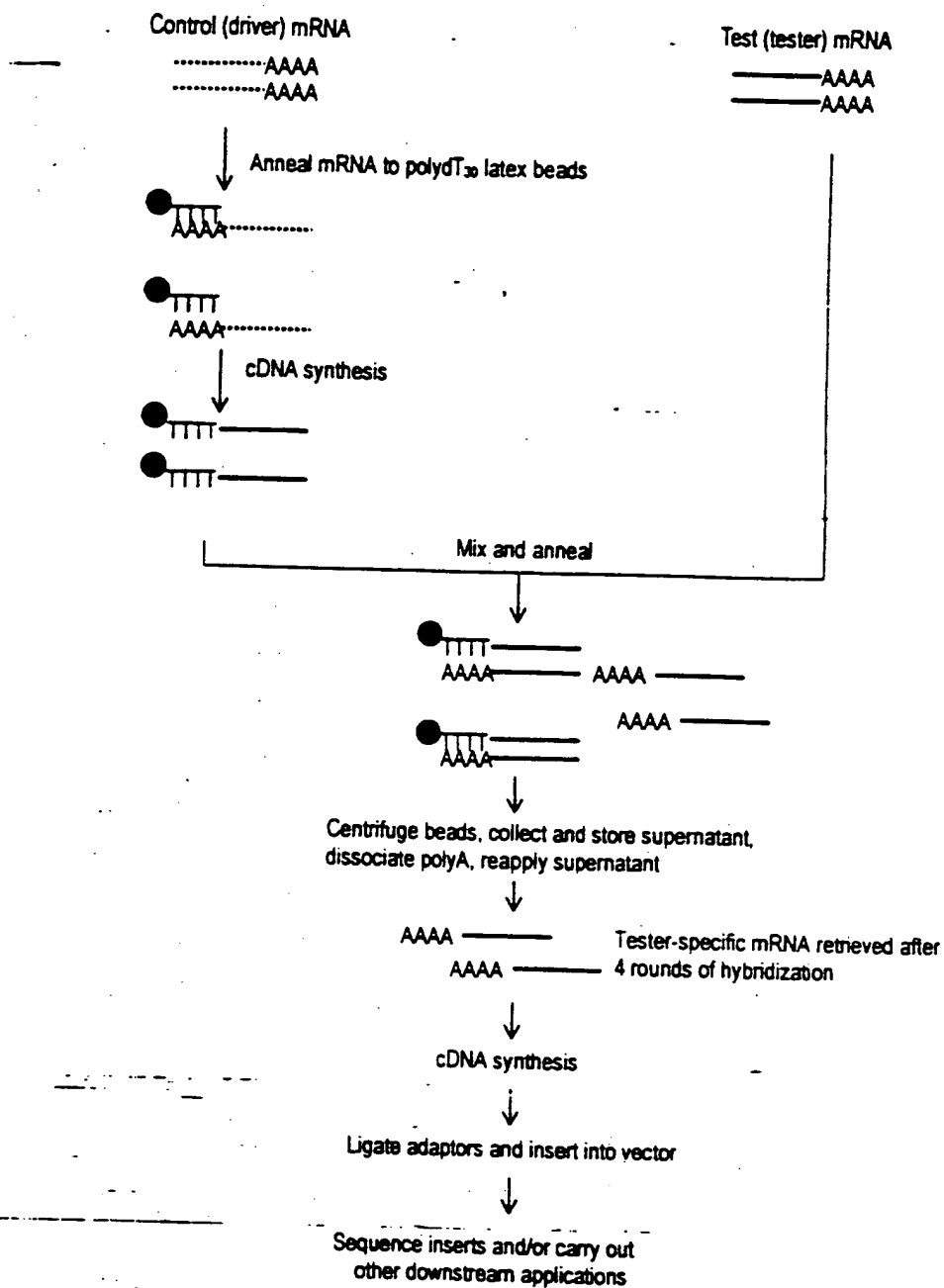


Figure 2. The use of oligodT<sub>30</sub> latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/alterd (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara et al. (1991).

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT<sub>30</sub>) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT<sub>30</sub> forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT<sub>30</sub> population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

#### Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 µg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promoter sequence

mRNA  
AA  
AA

retrieved after

A extracted from the  
dT oligonucleotides  
lation is repeatedly  
ulation of mRNA is  
ream applications, as

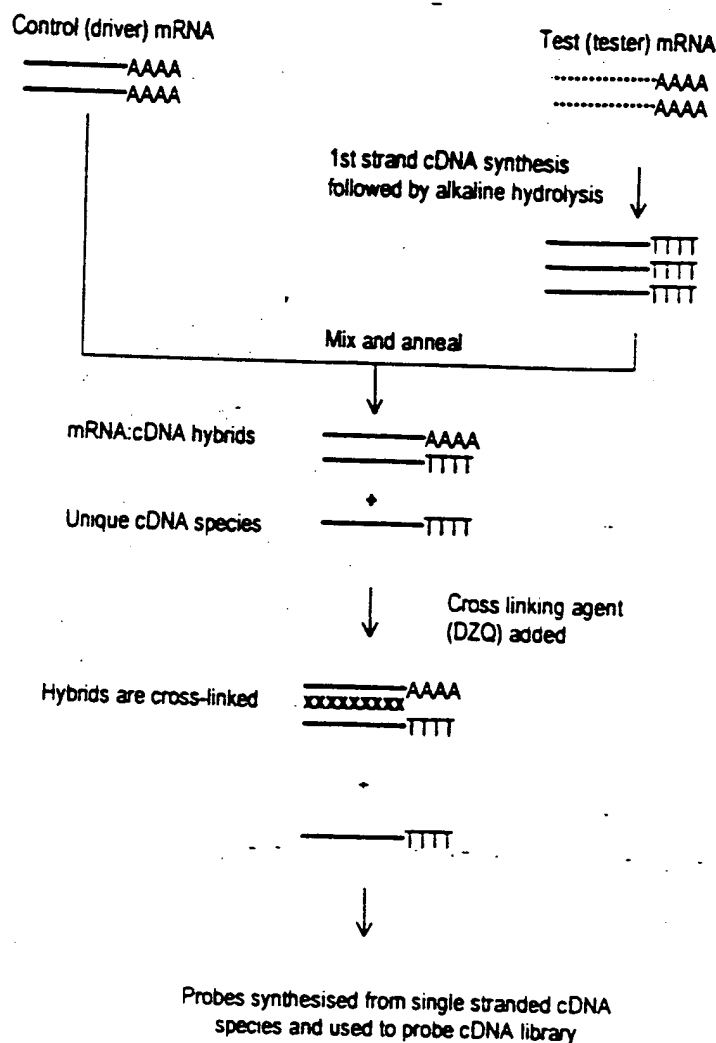


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1<sup>st</sup> strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,5 diazidiny-1,4-benzoquinone (DZO) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/ $\mu$ g total RNA
Abundant	12000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11000	0.004	0.002

— Modified from Bertoli *et al.* (1995).

mRNA

AAAA

AAAA

TTTT

TTTT

TTTT

at the 5' end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

#### Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5' adaptors at each end of the DNA duplex and can, thus, be filled in at both 3' ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Schatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

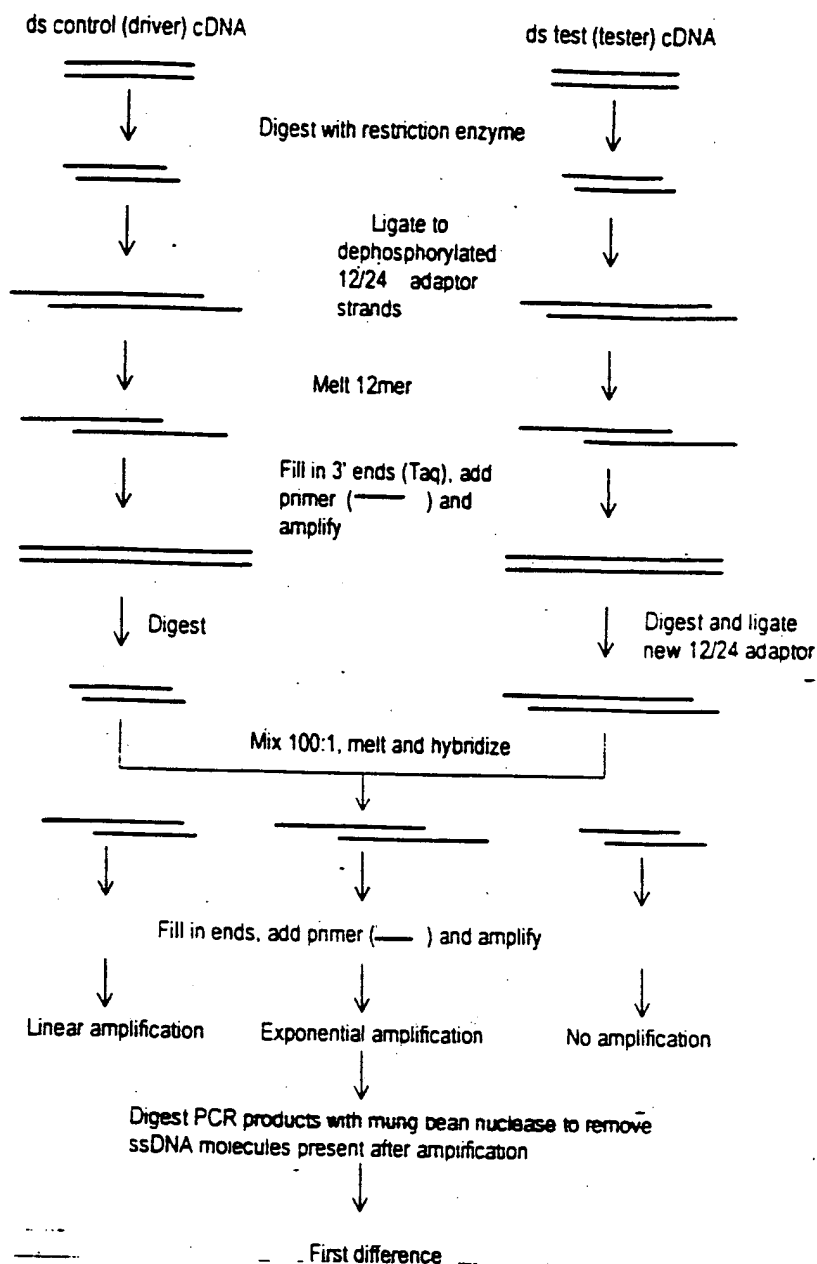
with 1<sup>st</sup> strand tester  
cross linked with 2.5  
ces are differentially  
using Sequenase 2.0  
does not react with the  
used to screen a cDNA  
ter *et al.* (1996), with

nalian cell.

1 mass  
of each  
es/ $\mu$ g  
RNA

65  
04  
002





**Figure 4.** The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *DpnII*. The 1<sup>st</sup> set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3' ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1<sup>st</sup> set of adaptors is removed with *DpnII*. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3' ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3<sup>rd</sup> or 4<sup>th</sup> difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

### Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

NA

st and ligate  
12/24 adaptor

on

and tester cDNA are  
12/24 adaptor strands  
products. The 12mer is  
polymerase. Each cDNA  
adaptor is removed with  
amplified tester cDNA  
of driver. The 12mer  
with primers identical  
which are exponentially  
ssDNA products are  
This is digested and a  
from the hybridization  
identified by Lisitsyn *et al.*

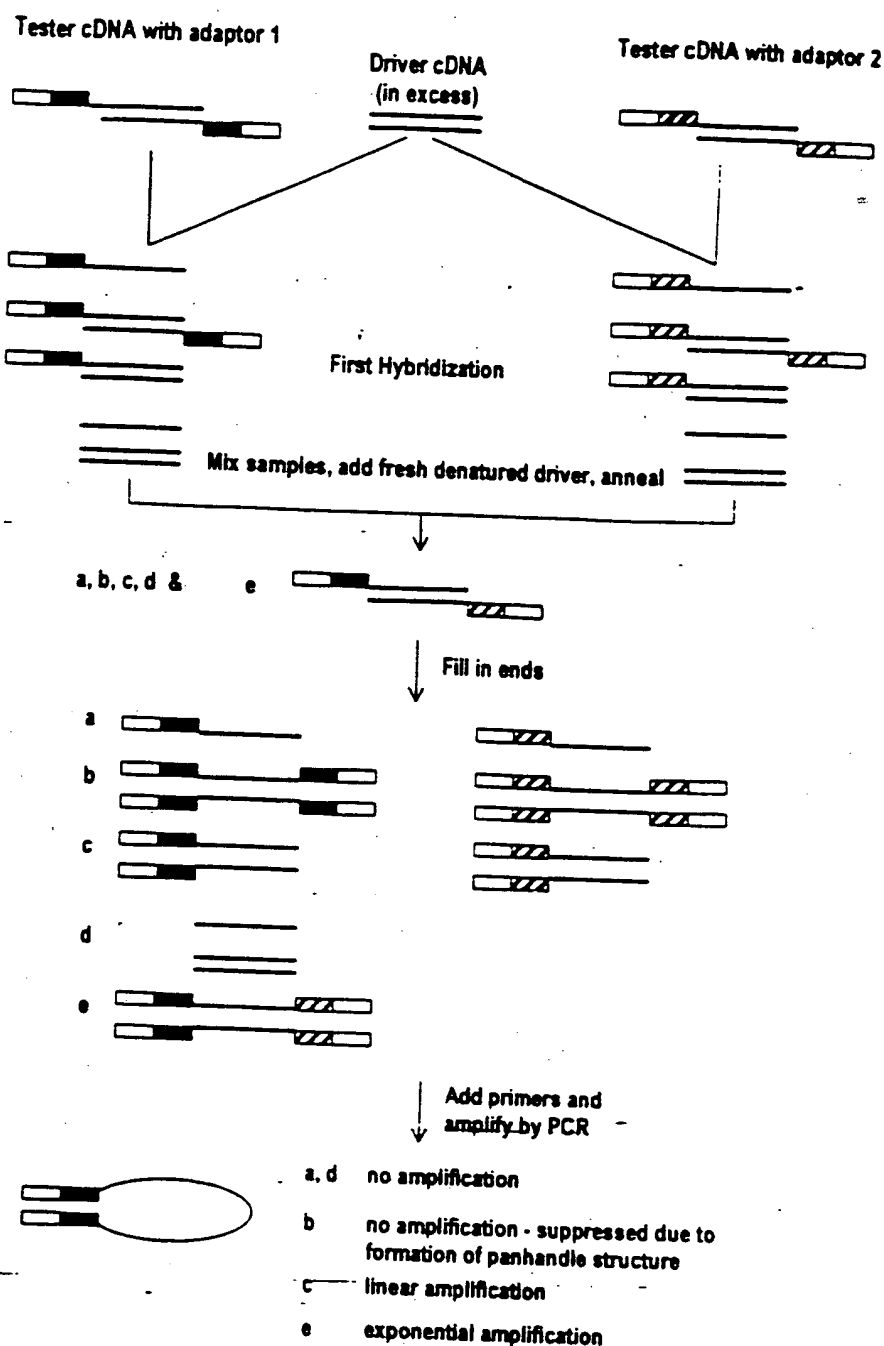


Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurakaya *et al.* (1996), with permission.

DNA with adaptor, 2

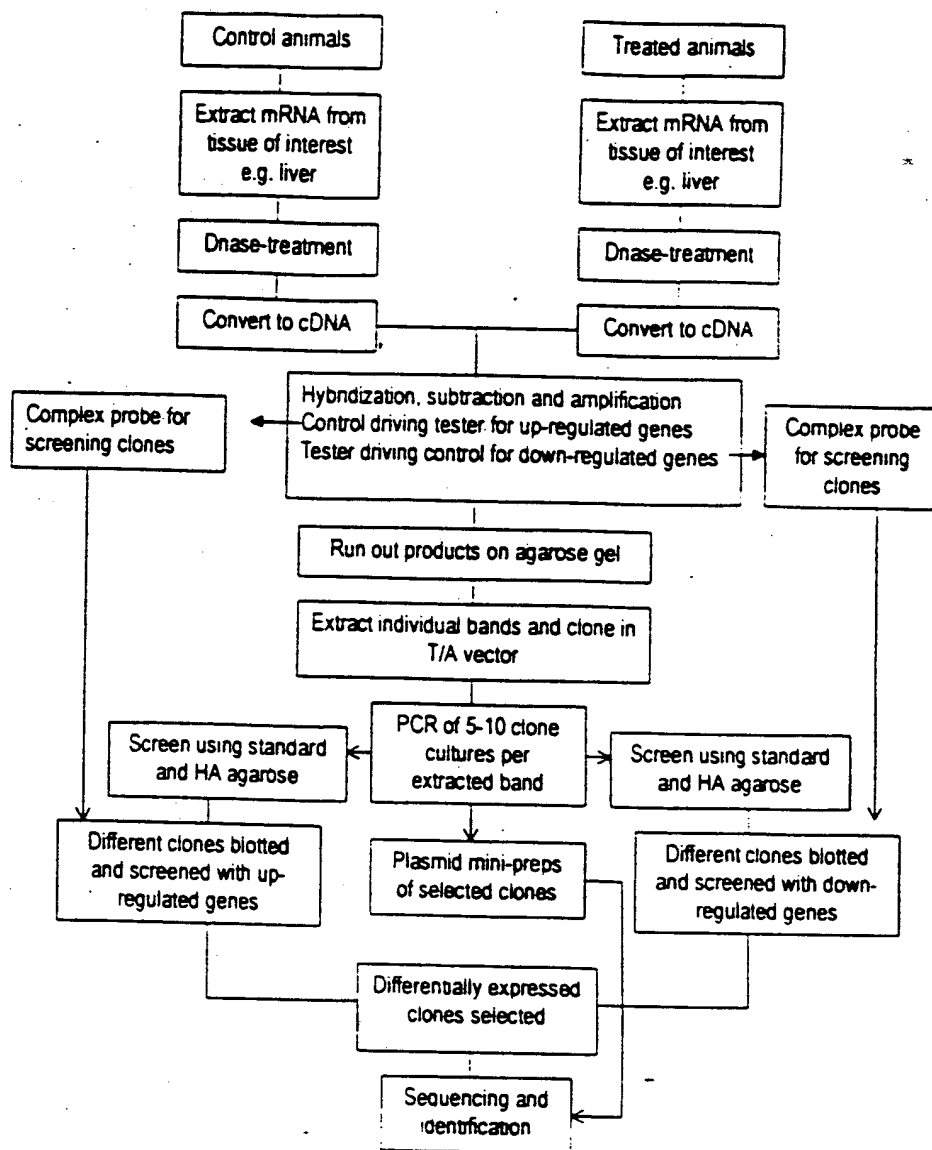
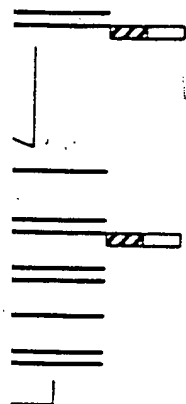


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3). One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

cess of driver cDNA is and allowed to hybridize abundant molecules; and differentially expressed tion, the two primary driver can also be added nces. Type e molecules fied using two rounds of irectly or cloned into a l. (1996) and Gurskaya

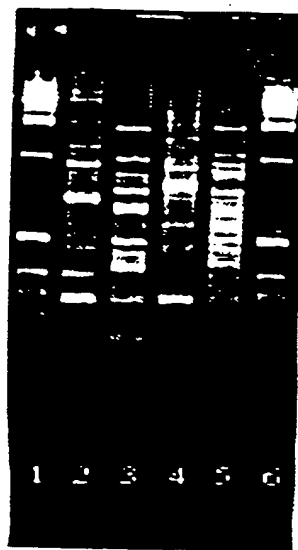


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with Wy-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy-14,643 treatment; 3—genes downregulated following Wy-14,643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al.* (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy-14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al.* 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete functional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

#### Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNA

Bands 1-4, 6, 9, 13, 14, and 17-20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopoxin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	<i>M. musculus</i> C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	<i>M. musculus</i> Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2 82.0%	Soares p3NMF 19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4-6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in on display only, or which are much more intense in one

ent with WY-14,643 or  
s used to generate the  
ontech). Lane: 1—1kb  
downregulated following  
d treatment; 5—genes  
oduced from Rockett *et*

ined. For example,  
en Wy,14,643, up-  
the rat (a sensitive  
the guinea pig, a  
hed observations).  
n-regulated in the  
named TAPA-1) is  
number of cellular  
rentiation (Levy *et*  
in the phenomena  
is intriguing, and  
rentially regulated  
of this approach is  
ase sequences, but  
nes of completely  
assessment of the  
e lack of complete  
: profiling studies  
biotic challenge,  
further detailed

imed PCR' (Liang  
to as 'differential

display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5' (dT<sub>11</sub>)CA 3' (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (*arbitrary* primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

- (1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guireraes *et al.* 1995a).
- (2) The cDNAs obtained often only represent the extreme 3' end of the mRNA (often the 3'-untranslated region), although this may not always be the case (Guireraes *et al.* 1995a). Since the 3' end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.
- (3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70% of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cytoplasmic RNA rather than total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

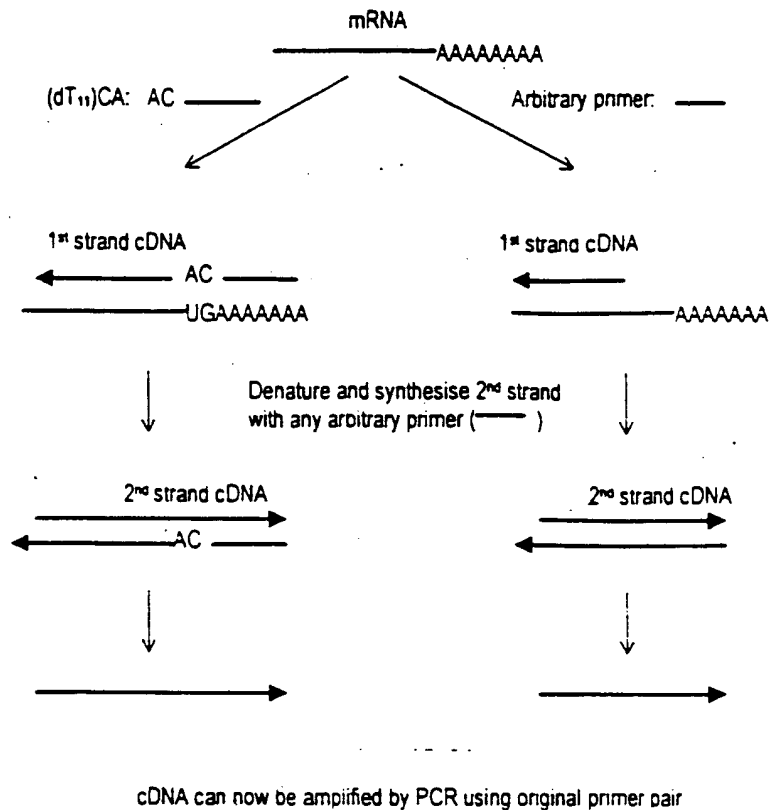


Figure 8. Two approaches to differential display (DD) analysis. 1<sup>st</sup> strand synthesis can be carried out either with a polydT<sub>11</sub>NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1<sup>st</sup> strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2<sup>nd</sup> strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2<sup>nd</sup> strand may also hybridize to the 1<sup>st</sup> strand cDNA in a number of different places, several different 2<sup>nd</sup> strand products may be obtained from one binding point of the 1<sup>st</sup> strand primer. Following 2<sup>nd</sup> strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

### Restriction endonuclease-facilitated analysis of gene expression

#### Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatenation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is



split into two and different adaptors ligated to the 5' ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

#### *Gene Expression Fingerprinting (GEF)*

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5' digestion products. The use of restricted 3'-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3' cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobilized 3'-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

group. Incorporated  
action enzyme—one  
ognition sequence.  
with the IIS enzyme,  
released. The two  
plified products are  
mers are formed in  
ndreds of gene tags  
the number of times  
ent of that gene's  
es identification of

al difficulty of the  
d towards abundant  
nic setting and has  
.

ating differentially  
ky (1995). In this  
(dT) primers. The  
and captured with  
wanted 5' digestion  
e complexity of the  
s is represented by  
cilitate subsequent  
with one adaptor-  
fied population is  
e dissociation. The  
it adapt r-specific  
obilized 3' cDNA  
tion endonucleases  
suit is a fingerprint  
nial digests used).  
entify differentially  
and cloned. The  
roducible, and the  
olved in the final  
arely resolve more  
more which are  
f 2-D gels such as  
1991) may help to

agments was later  
ead of s sequential  
se authors simply  
s without further

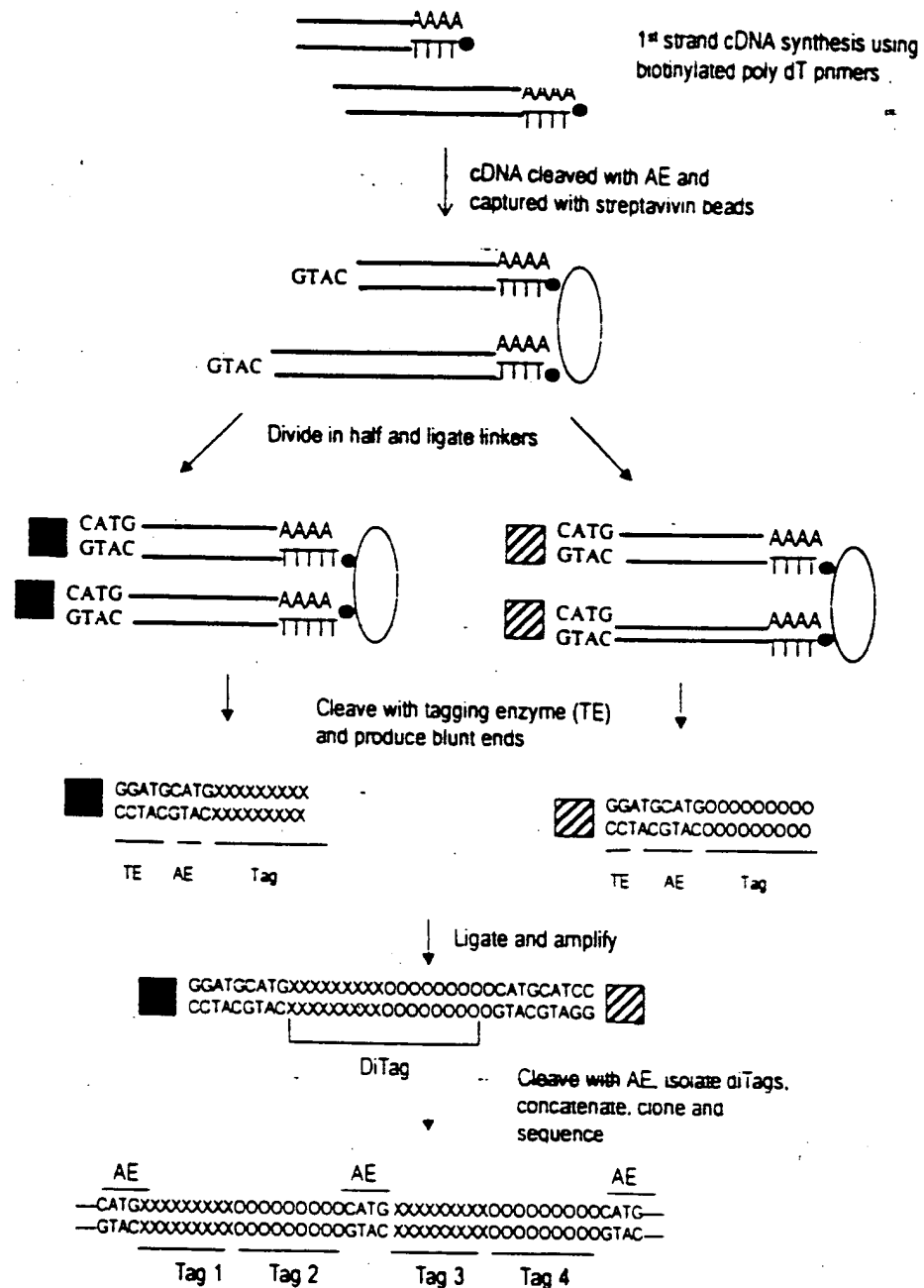


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3' ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the diTags isolated from the linkers using PAGE. The diTags are then ligated (during which process, concatenization occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

### DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

### EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmataz *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at <http://www.tigr.org>) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmataz *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

### Problems and potential of differential expression techniques

#### *The holistic or single cell approach?*

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdissection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: <http://www.ncbi.nlm.nih.gov/ncicgap/intro.html>). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,

it takes a great deal that they are indeed sue. Normally, the R. Even so, each of al of rapid analysis he development of ichena *et al.* 1996), ial gene expression 'chips' containing le copies of part of roven involvement cellular processes. und animal species. d a few companies a Laboratories and or even thousands NA from the test en analysed with ntitative means to s. Of course, there h are in the array o elucidating the ent system may be ectly identify and ons, and an open tially expressed.

of gene fragments ed gridding up to igh density chip- uced off-the-shelf determination of. Aside from their nd probing DNA ewer micro-array oducible between solved within the

#### sed g nes

es obtained from dentity (putative apid and efficient profiles of gene- ams *et al.* (1991), ted that there are esenting over half

e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromised tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth. It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

*How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?*

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15 000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30 000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005% of the total population (table 1). Bertoli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA

1998, Kas-Deelen *et al.* 1998).  
 issue unimportant.  
 altered expression  
 After all, since all  
 which intimately  
 h cell type could in  
 ecular mechanisms  
 owth. It is perhaps  
 nts using *in vivo* as  
 ical cells probably  
 ecular changes that

biological variation  
 s are being used. It  
 it ways to identical  
 soquine oxidation  
 ad determines the  
 d 1993, Meyer and  
 omplex, but allelic  
 and mental health  
 s. Careful thought  
 and to the possible  
 ct of this can be  
 nimportant minor  
 al animals, thus  
 echanisms of the  
 ay be of utmost  
 mb to or resist the

high percentage of

esting that mam-  
 es at any one time  
 lthough figures as  
 drick *et al.* (1984)  
 he rare abundance  
 ble 1.

en compared with  
 ot all diff rentially  
 lar, rare messages  
 easily r covered  
 as the majority of  
 ulation (table 1).  
 s (heterogeneous  
 : to detect mRNA

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000 × smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon- $\gamma$ -stimulated HeLa cells differentially express up to 433 genes (assuming 24000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate: Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3' restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3' polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle et al. 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

#### *How sensitive are differential expression technologies?*

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in expression of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between

effective—proving members of artificial are messages already models will genuinely in addition, there are , mRNAs may have plification by PCR-rcumstances not all lopment, deadenyl-iteitz 1998), whilst p70 (and perhaps, lle *et al.* 1994). The ficiency of systems of any system also display techniques olate mRNA that is used to prime first ed to some degree has been shown, at a lead to inefficient braction kit user ewise in other SH n amplification step sequences amplify

temporal factor. It interrogate a cell at nes showing altered ease processes and des of signalling, which are switched al information may ormation about the can be derived for lar interest to the ne point analysis is h, of course, adds

of how large the re in qu stion with isolation of genes ported using SSH rating a change in e is a 'grey zon' isolation between

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

#### *Resolution and visualization of differential expression products*

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hessele, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hans Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot



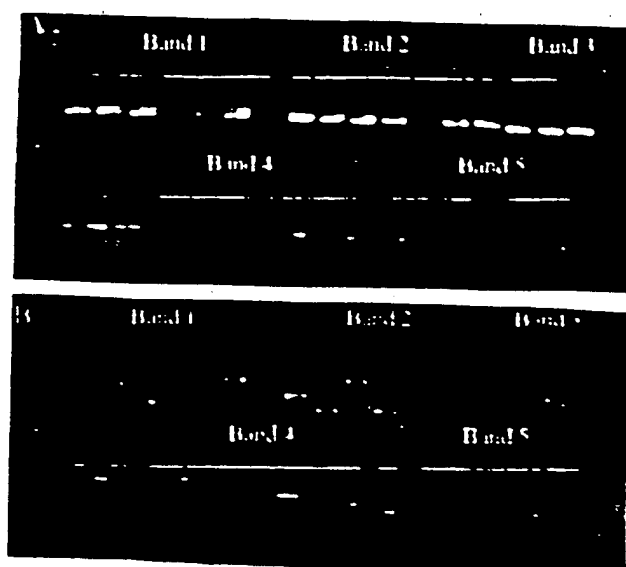


Figure 10. Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2% agarose gel, and (B) a high resolution 2% agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, over staining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to over stain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

### The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000+ bp on the same gel, which leads to compromise in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to include certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

red. Bands of decreasing  
attractive hybridization  
h cloned band and their  
h resolution 2% agarose  
With few exceptions, all  
the presence of HA-red  
percentage of GC within  
within each band. For  
ie same size, at least four

gel should indicate  
separate otherwise  
content. Geisinger  
ifying DD-derived  
laboratory on clones

differential display  
d out in a standard  
d and incorporated

ere being different  
content. However,  
in, one might use  
GE) or temperature  
contents f a band,  
on the reamplified

to visualize large  
m in that, in terms  
s. One approach to  
by Uitterlind n et

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

### Screening

#### *False positives*

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitimate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

### Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene  $X_0$  really come from that gene, or its brother gene  $X_1$  or its as yet undiscovered sister  $X_2$ ? For example, using SSH, part of a gene was isolated,

ne altered expression primers and/or post-receptors, cell cycling is considered as candidates rays (e.g. Clontech's is to some degree by apoptosis, stress, DNA-

length amongst the *et al.* 1994, Sun *et al.* varies with the factors which have not through illegitimate can arise through also positives appear some may arise from technical reasons. can be carried out probes synthesized clones (Hedrick *et al.* will hybridize to is that rare species or those using SSH the subtracted cDNA reverse subtraction has rare sequences. ting low abundance d to go back to the more quantitative is, the sensitivity is ds for accurate and

products which are y reduce the size of leads to a reduced bers whose DNA g. the cytochrome identified as being other gene  $X_1$  or its gene was isolated,

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

- (1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
- (2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

### Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using multichannel pipettes, 96+-well plates and

appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN- $\gamma$ , Frye *et al.* 1989),  $\beta$ -actin (Heuval *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991),  $\beta$ -2-microglobulin ( $\beta$ -2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al.* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

#### **Problems in using the differential display approach**

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has

e analysis is more internal standard, the is oft n excessive, of gene species. The ively involved. One age in the test cells tried in the past, for Heuval *et al.* 1994), *et al.* 1994), di-microglobulin ( $\beta$ -2-use (HPRT, Foss *et* Ideally, an internal gardless of cell age, however, it has been ing genes currently n conditions and in herefore, that pre-g genes to establish

with caution. By n one can perhaps to external stimuli. ic effects of a wide nea pigs are largely *et al.* 1989, 1993, reason(s) why is to ify those which are lge of the effects of ox ic carcinogenesis omplex. Perhaps if c effects and it was egulated five times imp rtance of the ange in expression ample, what is the rticular treatment, the literature one -regulated 40-60- old increase would e Z has never been nakes your 5-fold if that same 5-fold tment with related

easily obtainable st animals/cells in . However, it has

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

### Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5-10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as described extensively herein, but rather protein-protein, protein-DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

#### Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

not hope to answer the question of the molecular mechanism of regulation and function of a fixed moment in time and the placement of the gene is asked to try and find the lid from a few still in the battle, the Historian and their commanding officer must examine the prevailing weather. As the forthcoming, the techniques, such as microarray analysis and Northern blotting, as emphasized the need for isolation and in combination with Northern blots from isoelectric focusing maps using capillary electrophoresis as many of the changes in mRNA levels, protein-DNA and functional genomics or

it is clear that many of the characterizing the genetic content and in response to such profiling will be used, and in the long run markers for different potential medical and diseases are almost indicated the family or to plus the length of prudent treatment. In all cancers, provide perhaps indicate the

next century and the ongoing development will ensure that this disease processes.

y Darney and Chris manuscript prior to with the policy of the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

## References

- ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMERPOULOS, M. H., XIAO, H., MERRILL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. and VENTOR, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651-1656.
- AN, G., LUO, G., VELTRI, R. W. and O'HARA, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342-346.
- AXEL, R., FEIGELSON, P. and SCHULTZ, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247-254.
- BAND, V. and SAGER, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249-1253.
- BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. and STRAUSS, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272-4280.
- BERTIOLI, D. J., SCHLICHTER, U. H. A., ADAMS, M. J., BURROWS, P. R., STEINBISS, H.-H. and ANTONIW, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520-4523.
- BRAVO, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37-46.
- BURN, T. C., PETROVICK, M. S., HOMAUS, S., ROLLINS, B. J. and TENEN, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776-2783.
- CAO, J., CAI, X., ZHENG, L., GENG, L., SHI, Z., PAO, C. C. and ZHENG, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447-451.
- CASSIDY, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13-20.
- CHANG, G. W. and TERZAGHI-HOWE, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445-4452.
- CHEN, J., SCHWARTZ, D. A., YOUNG, T. A., NORRIS, J. S. and YAGER, J. D., 1996, Identification of genes whose expression is altered during mitosis suppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783-2786.
- CHEN, J. J. W. and PECK, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793-794.
- CLONTECHNIQUES, 1997a, PCR-Select Differential Screening Kit—the next step after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18-19.
- CLONTECHNIQUES, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15-16.
- DAVIS, M. M., COHEN, D. I., NIELSEN, E. A., STEINMETZ, M., PAUL, W. E. and HOOD, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194-2198.
- DELLAVALLE, R. P., PETERSON, R. and LINDQUIST, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in *Drosophila melanogaster*. *Molecular and Cell Biology*, **14**, 3646-3659.
- DE RISI, J. L., VASHWANATH, R. L. and BROWN, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.
- DIATCHENKO, L., LAU, Y.-F. C., CAMPBELL, A. P., CHENCHIK, A., MOQADAM, F., HUANG, B., LUKYANOV, K., GURSKAYA, N., SVERDLOV, E. D. and SIEBERT, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025-6030.
- DOGRA, S. C., WHITELAW, M. L. and MAY, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1-9.
- DUGUID, J. R. and DINAUER, M. C., 1990, Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789-2792.
- DUNBAR, P. R., OGG, G. S., CHEN, J., RUST, N., VAN DER BRUGGEN, P. and CERUNDOLLO, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413-416.



- FITZPATRICK, D. R., GERMAIN-LEE, E. and VALLE, D., 1995, Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, **27**, 457-466.
- FOSS, D. L., BAARSCH, M. J. and MURTAUGH, M. P., 1998, Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, **9**, 67-78.
- FRYE, R. A., BENZ, C. C. and LIU, E., 1989, Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, **4**, 1153-1157.
- GEISINGER, A., RODRIGUEZ, R., ROMERO, V. and WETTSTEIN R., 1997, A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, <http://tto.trends.com>, document T01110.
- GRESS, T. M., HOHEISEL, J. D., LENNON, G. G., ZEHETNER, G. and LEHRACH, H., 1992, Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, **3**, 609-619.
- GRIFFIN, G. and KRISHNA, S., 1998, Cytokines in infectious diseases. *Journal of the Royal College of Physicians, London*, **32**, 195-198.
- GROENINK, M. and LEEGWATER, A. C. J., 1996, Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechniques*, **XI**, 23-24.
- GUIMARAES, M. J., BAZAN, J. F., ZLOTNIK, A., WILES, M. V., GRIMALDI, J. C., LEE, F. and McCLANAHAN, T., 1995b, A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, **121**, 3335-3346.
- GUIMARAES, M. J., LEE, F., ZLOTNIK, A. and McCLANAHAN, T., 1995a, Differential display by PCR: novel findings and applications. *Nucleic Acids Research*, **23**, 1832-1833.
- GURSKAYA, N. G., DIATCHENKO, L., CHENCHIK, P. D., SIEBERT, P. D., KHASPEKOV, G. L., LUKYANOV, K. A., VAGNER, L. L., ERMOLAEVA, O. D., LUKYANOV, S. A. and SVERDLOV, E. D., 1996, Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-Myristate 13-Acetate. *Analytical Biochemistry*, **240**, 90-97.
- HAMPSON, I. N. and HAMPSON, L., 1997, CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), **23**, 22-24.
- HAMPSON, I. N., HAMPSON, L. and DEXTER, T. M., 1996, Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*, **24**, 4832-4835.
- HAMPSON, I. N., POPE, L., COWLING, G. J. and DEXTER, T. M., 1992, Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, **20**, 2899.
- HARA, E., KATO, T., NAKADA, S., SEKIYA, S. and ODA, K., 1991, Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, **19**, 7097-7104.
- HATADA, I., HAYASHIZAKE, Y., HIROTSUNE, S., KOMATSUBARA, H. and MUKAI, T., 1991, A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences (USA)*, **88**, 9523-9527.
- HECHT, N., 1998, Molecular mechanisms of male sperm cell differentiation. *Bioessays*, **20**, 555-561.
- HEDRICK, S., COHEN, D. I., NIELSEN, E. A. and DAVIS, M. E., 1984, Isolation of T cell-specific membrane-associated proteins. *Nature*, **308**, 149-153.
- HERTZ, R., SECKBRACH, M., ZAKIN, M. M. and BAR-TANA, J., 1996, Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, **271**, 218-224.
- HEUVAL, J. P. V., CLARK, G. C., KOHN, M. C., TRITSCHER, A. M., GREENLEE, W. F., LUCIER, G. W. and BELL, D. A., 1994, Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transcriptase-polymerase chain reaction. *Cancer Research*, **54**, 62-68.
- HILLIER, L. D., LENNON, G., BECKER, M., BONALDO, M. F., CHIAPELLI, B., CHISSOE, S., DIETRICH, N., DUBOQUIE, T., FAYELLO, A., GISH, W., HAWKINS, M., HULTMAN, M., KUCABA, T., LACY, M., LE, M., LE, N., MARDIS, E., MOORE, B., MORRIS, M., PARSONS, J., PRANGE, C., RIFKIN, L., ROHLFING, T., SCHELLENBERG, K., SOARES, M. B., TAN, F., THIERRY-MEG, J., TREVASKIS, E., UNDERWOOD, K., WOHLDMAN, P., WATERSTON, R., WILSON, R. and MARRA, M., 1996, Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807-828.
- HUBANK, M. and SCHATZ, D. G., 1994, Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, **22**, 5640-5648.
- HUNTER, T., 1991, Cooperation between oncogenes. *Cell*, **64**, 249-270.
- IVANOVA, N. B. and BELYAVSKY, A. V., 1995, Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, **23**, 2954-2958.
- JAMES, B. D. and HIGGINS, S. J., 1985, *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).
- KAS-DEELEN, A. M., HARMSSEN, M. C., DE MAAR, E. F. and VAN SON, W. J., 1998, A sensitive method for

- characterisation of rat and A hydratase. *Genomics*, 27,
- of hypoxanthine phospho-eta-actin mRNA expression
- by differential polymerase
- simple method for screening is. *Elsevier Trends Journals*
- CH, H., 1992. Hybridisation derived from whole tissues.
- rnal of the Royal College of
- genes associated with liver techniques, XI, 23-24.
- ALDI, J. C., LEE, F. and etic development in the yolk
- ia, Differential display by 32-1833.
- USPEKOV, G. L., LUKYANOV, d SVERDLOV, E. D., 1996, polymerase chain reaction: l phorbol 12-Myristate 13-
- ing made easy. *Life Science*
- om oligonucleotide primed ve cDNA cloning. *Nucleic*
- cal cross linking subtraction tion probes. *Nucleic Acids*
- ctive cDNA cloning using o undifferentiated human
- KAI, T., 1991, A genomic dmarks. *Proceedings of the Bioessays*, 20, 555-561.
- isolation of T cell-specific
- optional suppression of the i *Biological Chemistry*, 271.
- W. F., LUCIER, G. W. and sponse relationships using research, 54, 62-68.
- CHISSOE, S., DIETRICH, N., KUCABA, T., LACY, M., LE, C., RIFKIN, L., ROHLFING, REVASKIS, E., UNDERWOOD, 6, Generation and analysis 828.
- ression by representational
- ressed genes by restriction earch, 23, 2954-2958.
- RL Press Ltd).
- 98, A sensitive method for quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, 5, 622-626.
- KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), 10, 50-51.
- KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, 7, 1611-1618.
- KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, 18, 5705-5711.
- LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, 101, 241-248.
- LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicology and Applied Pharmacology*, 99, 148-160.
- LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, 9, 60-77.
- LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, 16, 89-109.
- LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967-971.
- LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, 52, 6966-6968.
- LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, 21, 3269-3275.
- LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOF, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, 254, 304-321.
- LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, 23, 3244-3251.
- LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, 259, 946-951.
- LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, 18, 200-202.
- LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Technology*, 17, 84-94.
- MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepatic cytochrome P450A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, 7, 183-191.
- MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, 53, 39-51.
- MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, 24, 1504-1507.
- MCKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), 10, 19-20.
- MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, 11, 242-246.
- MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, 88, 29-36.
- MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, 37, 269-296.
- MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, 28, 437-447.
- MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TRITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the mdr-1 gene expression. *Biochemistry*, 29, 10351-10356.
- NELSON, D. R., KOYMAN, L., KAMATAKI, T., STEGEMAN, J. J., FEYERISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, 6, 1-42.

- NISHIO, Y., AIELLO, L. P. and KING, G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, **8**, 103-106.
- O'NEILL, M. J. and SINCLAIR, A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, **25**, 2681-2682.
- ORTON, T. C., ADAM, H. K., BENTLEY, M., HOLLOWAY, B. and TUCKER, M. J., 1984, Clobuzant: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, **73**, 138-151.
- PELKONEN, O., MAENPAA, J., TAAVITSAINEN, P., RAUTIO, A. and RAUNIO, H., 1998, Inhibition and Induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*, **28**, 1203-1253.
- PHILIPS, S. M., BENDALL, A. J. and RAMSHAW, I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, **82**, 199-203.
- PRASHAR, Y. and WEISSMAN, S. M., 1996, Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences (USA)*, **93**, 659-663.
- RAGNO, S., ESTRADA, I., BUTLER, R. and COLSTON, M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, **57**, 143-146.
- RAMANA, K. V. and KOHLI, K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, **36**, 437-446.
- RICHARD, L., VELASCO, P. and DETMAR, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, **240**, 1-6.
- ROCKETT, J. C., ESDAILE, D. J. and GIBSON, G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism and Pharmacokinetics*, **22**, 329-333.
- RODRICKS, J. V. and TURNBULL, D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, **3**, 197-212.
- ROGLER, G., HAUSMANN, M., VOGL, D., ASCHENBRENNER, E., ANDUS, T., FALK, W., ANDRESEN, R., SCHOLMERICH, J. and GROSS, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, **112**, 205-215.
- ROHN, W. M., LEE, Y. J. and BENVENISTE, E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, **16**, 311-330.
- RUDIN, C. M. and THOMPSON, C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, **25**, 435-446.
- SAKAGUCHI, N., BERGER, C. N. and MELCHERS, F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, **5**, 2139-2147.
- SAMBROOK, J., FRITSCH, E. F. and MANIATIS, T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Ferguson (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6-37.
- SARGENT, T. D. and DAWID, I. B., 1983, Differential gene expression in the gastrula of *Xenopus laevis*. *Science*, **222**, 135-139.
- SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O. and DAVIS, R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences (USA)*, **93**, 10614-10619.
- SCHNEIDER, C., KING, R. M. and PHILIPSON, L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, **54**, 787-793.
- SCHNEIDER-MALNOURY, S., GILARDI-HEBENSTREIT, P. and CHARNAY, P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, **321**, 819-834.
- SEMENTA, G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, **3**, 180-199.
- SEWALL, C. H., BELL, D. A., CLARK, G. C., TRITSCHER, A. M., TULLY, D. B., VANDEN HEUVEL, J. and LUCIER, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, **41**, 1829-1834.
- SINGH, N., AGRAWAL, S. and RASTOGI, A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, **3**, 41-49.
- SMITH, N. R., LI, A., ALDERSLEY, M., HIGH, A. S., MARKHAM, A. F. and ROBINSON, P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids Research*, **25**, 3552-3554.
- SOMPAYRAC, L., JANE, S., BURN, T. C., TENEN, D. G. and DANNA, K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, **23**, 4738-4739.
- ST JOHN, T. P. and DAVIS, R. W., 1979, Isolation of galactose-inducible DNA sequences from *Saccharomyces cerevisiae* by differential plaque filter hybridisation. *Cell*, **16**, 443-452.
- SUN, Y., HEGAMYER, G. and COLBURN, N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, **54**, 1139-1144.

- ovine aortic smooth muscle 3-106.  
representational difference
- .. 1984, Clobazart: species e liver following chronic
- H., 1998, Inhibition and 8, 1203-1253.  
enes associated with high National Cancer Institute.
- ession by display of 3' end y of Sciences (U.S.A.) 93,
- tion of macrophage gene gy Letters, 57, 143-146.  
450—an overview. Indian
- protocol for the selective thelial cells. *Experimental*
- ofiling of non-genotoxic olymerase chain reaction etics, 22, 329-333.  
xisomes and peroxisome
- ALK, W., ANDRESEN, R., aracterization of colonic
- AHC expression. *Critical*
- n. *Seminars in Oncology*,
- copy of an RNA species
- f DNA. In N. Ford, M. 2nd edition (New York:
- strula of *Xenopus laevis*.
- V., 1996, Parallel human nes. *Proceedings of the*
- essed at growth arrest of
- low to build a vertebrate -834.  
ns and pathophysiology.
- VANDEN HEUVEL, J. and r biomarkers. *Clinical*
- unity: special reference 49.
- SON, P. A., 1997, Rapid F-PCR polyacrylamide
- Overcoming limitations 4738-4739.  
DNA sequences from 16, 443-452.  
five messenger RNAs idermal cells: one is search, 54, 1139-1144.
- SUNG, Y. J. and DENMAN, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, 23, 462-464.
- SUTTON, G., WHITE, O., ADAMS, M. and KERLAVAGE, A., 1995, TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1, 9-19.
- SUZUKI, Y., SEKIYA, T. and HAYASHI, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, 192, 82-84.
- SYED, V., GU, W. and HECHT, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, 18, 264-273.
- UTTERLINDEN, A. G., SLAGBOOM, P., KNOOK, D. L. and VIJGL, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (U.S.A.)*, 86, 2742-2746.
- ULLMAN, K. S., NORTHROP, J. P., VERWEIJ, C. L. and CRABTREE, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, 8, 421-452.
- VASMATZIS, G., ESSAND, M., BRINKMANN, U., LEE, B. and PASTON, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (U.S.A.)*, 95, 300-304.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W., 1995, Serial analysis of gene expression. *Science*, 270, 484-487.
- VOELTZ, G. K. and STEITZ, J. A., 1998, AUAUA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. *Molecular and Cell Biology*, 18, 7537-7545.
- VOGELSTEIN, B. and KINZLER, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, 9, 138-141.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, 21, 13-14.
- WAN, J. S., SHARP, S. J., POIRIER, G. M.-C., WAGAMAN, P. C., CHAMBERS, J., PYATI, J., HOM, Y.-L., GALINDO, J. E., HUVAR, A., PETERSON, P. A., JACKSON, M. R. and ERLANDER, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, 14, 1685-1691.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, 21, 13-14.
- WANG, Z. and BROWN, D. D., 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (U.S.A.)*, 88, 11505-11509.
- WAWER, C., RUGGEBERG, H., MEYER, G. and MUYZER, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, 23, 4928-4929.
- WELSH, J., CHADA, K., DALAL, S. S., CHENG, R., RALPH, D. and MCCLELLAND, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, 20, 4965-4970.
- WONG, H., ANDERSON, W. D., CHENG, T. and RIABOWOL, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, 223, 251-258.
- WONG, K. K. and MCCLELLAND, M., 1994, Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (U.S.A.)*, 91, 639-643.
- WYNFORD-THOMAS, D., 1991, Oncogenes and anti-oncogenes: the molecular basis of tumour behaviour. *Journal of Pathology*, 165, 187-201.
- XHU, D., CHAN, W. L., LEUNG, B. P., HUANG, F. P., WHEELER, R., PIEDRAFITA, D., ROBINSON, J. H. and LIEW, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, 187, 787-794.
- YANG, M. and SYTOWSKI, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, 237, 109-114.
- ZHAO, N., HASHIDA, H., TAKAHASHI, N., MISUMI, Y. and SAKAKI, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, 156, 207-213.
- ZHAO, X. J., NEWSOME, J. T. and CIHLAR, R. L., 1998, Up-regulation of two *Candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, 25, 121-129.
- ZIMMERMANN, C. R., ORR, W. C., LECLERC, R. F., BARNARD, C. and TIMBERLAKE, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, 21, 709-715.

## IN PERSPECTIVE

Claudio J. Conti, Editor

# Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,<sup>1</sup> Michael Bittner,<sup>2</sup> Jeffrey Trent,<sup>2</sup> J. Carl Barrett,<sup>1</sup> and Cynthia A. Afshari<sup>1</sup><sup>1</sup>Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina<sup>2</sup>Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153-159, 1999. © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

## INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10-12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

## MICROARRAY DEVELOPMENT AND APPLICATIONS

### cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

\*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorophores. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22-24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25-27].

#### Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28-30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)<sup>+</sup> RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

### THE USE OF MICROARRAYS IN TOXICOLOGY

#### Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

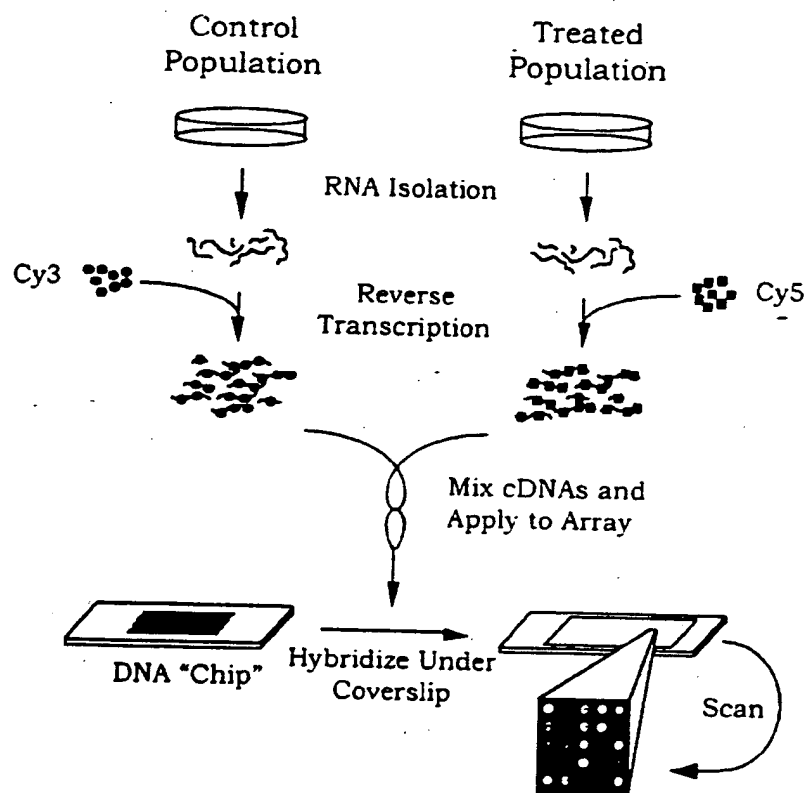


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

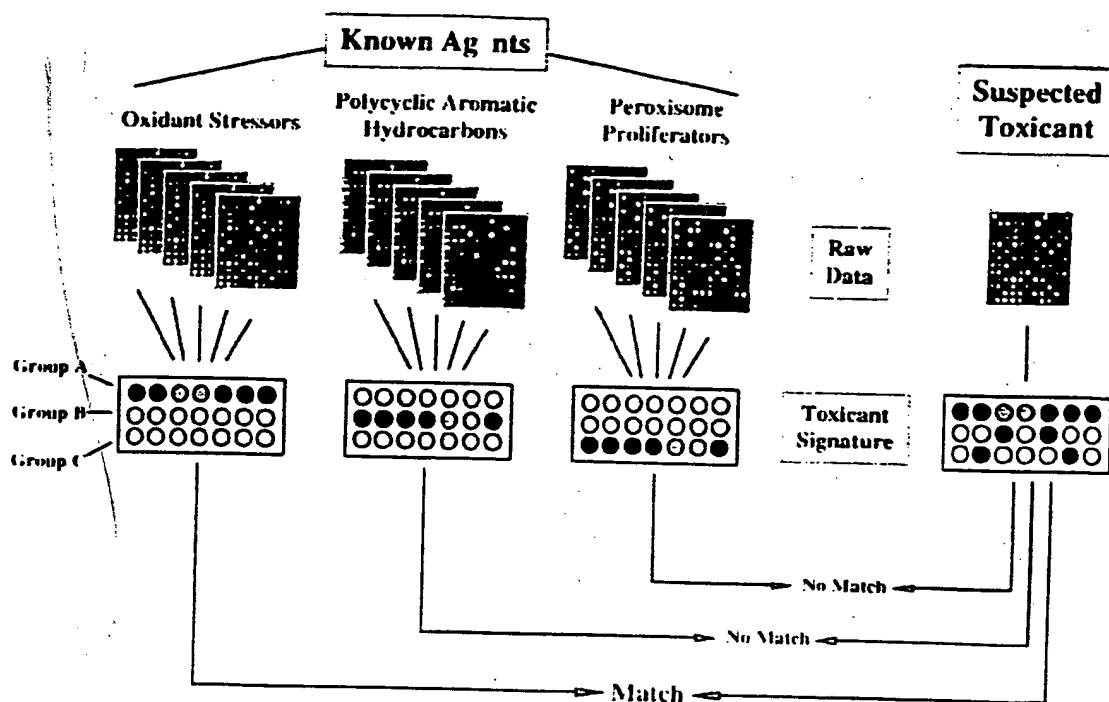


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing Tox-Chip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

#### Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown



Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Design d to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

\*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

#### Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

### Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

### FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

### ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

### REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996;93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nngri.nih.gov/DIR/LCG/15K/HTML/dabase.htm>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>



Toxicology Letters 112–113 (2000) 467–471

Docket No.: PF-0505-2 DIV

ISSN: 0978-2843

Enclosure No. 4 of 5

## Toxicology Letters

[www.elsevier.com/locate/toxlet](http://www.elsevier.com/locate/toxlet)

# Expression profiling in toxicology — potentials and limitations

Sandra Steiner \*, N. Leigh Anderson

*Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA*

### Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

**Keywords:** Proteomics; Genomics; Toxicology

### 1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the

pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

\* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: [steiner@lsbc.com](mailto:steiner@lsbc.com) (S. Steiner)

## 2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

## 3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

## 4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

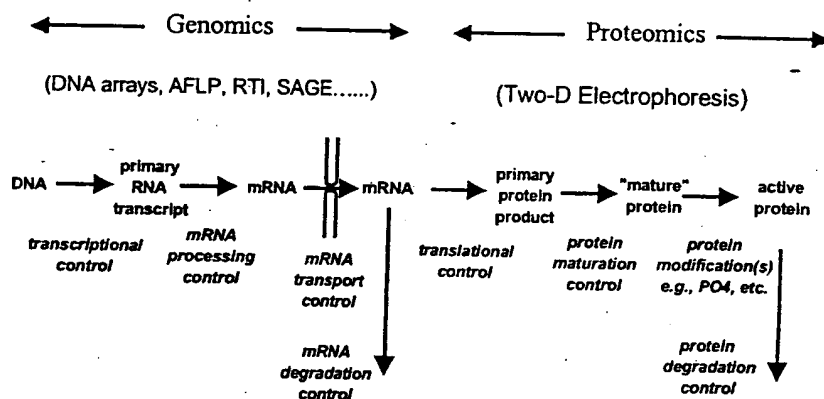


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

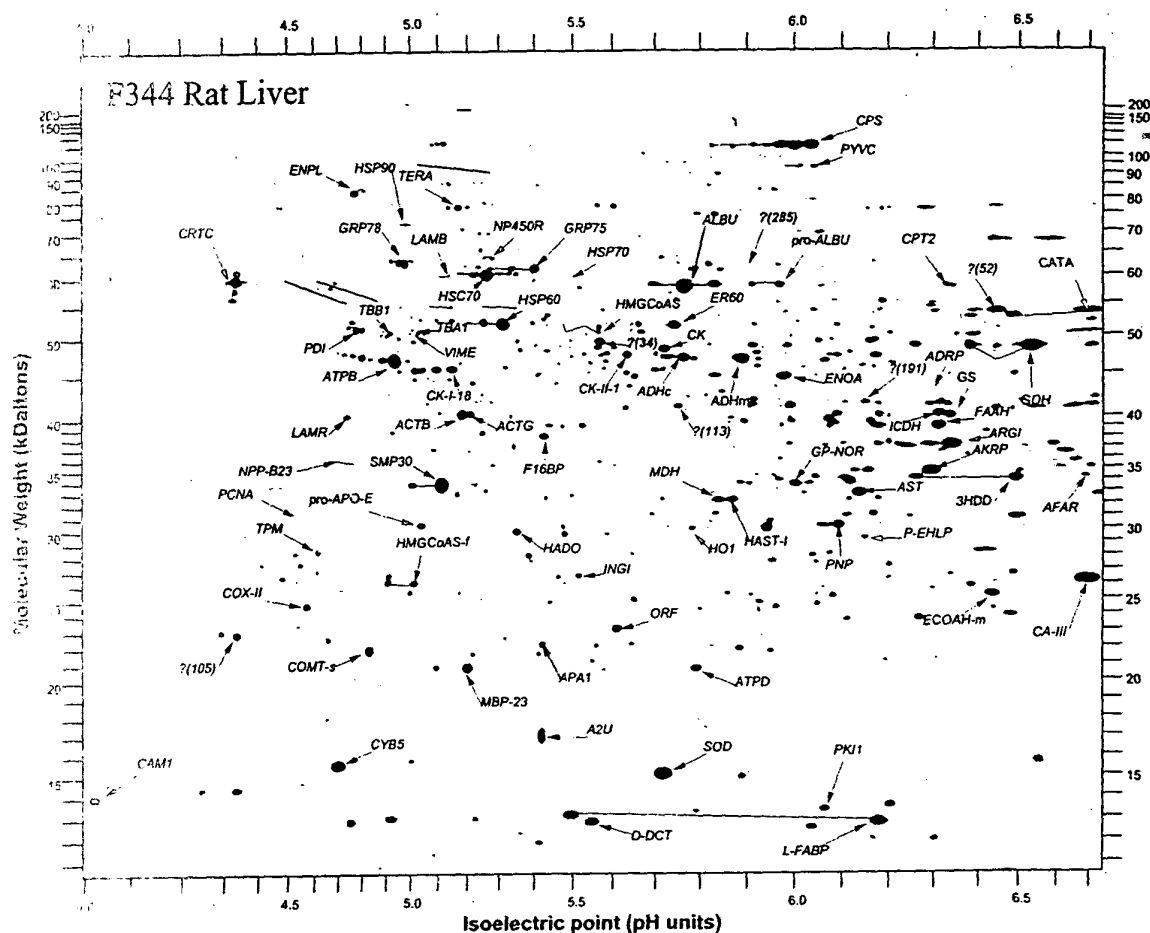


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate

### 5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

## 6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

## 7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trials.

## References

- Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19, 1998–2003.
- Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12, 907–930.
- Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell. Biochem. Suppl.* 22, 108–116.
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63, 2243–2250.

- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. *Science* 274, 610–614.
- Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. *Electrophoresis* 19, 355–363.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773.
- Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345.
- Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. *Electrophoresis* 14, 157–161.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 251, 467–470.
- Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645.
- Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raynackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir-treated rats. *Biochem. Biophys. Res. Commun.* 218, 777–782.
- Steiner, S., Aicher, L., Raynackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. *Biochem. Pharmacol.* 51, 253–258.
- Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. *Curr. Biol.* 6, 1543–1544.



Subject: RE: [Fwd: Toxicology Chip]

Date: Mon. 3 Jul 2000 08:09:45 -0400

From: "Afshari, Cynthia" <afshari@niehs.nih.gov>

To: "Diana Hamlet-Cox" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at  
<http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox  
response and basic cellular processes and added a set of clones and ESTs to  
this. We have included a set of control genes (80+) that were selected by  
the NHGRI because they did not change across a large set of array  
experiments. However, we have found that some of these genes change  
significantly after tox treatments and are in the process of looking at the  
variation of each of these 80+ genes across our experiments.  
Our chips are constantly changing and being updated and we hope that our  
data will lead us to what the toxchip should really be.  
I hope this answers your question.

Cindy Afshari

> -----  
> From: Diana Hamlet-Cox  
> Sent: Monday, June 26, 2000 8:52 PM  
> To: afshari@niehs.nih.gov  
> Subject: [Fwd: Toxicology Chip]

> Dear Dr. Afshari,

> Since I have not yet had a response from Bill Grigg, perhaps he was not  
> the right person to contact.

> Can you help me in this matter? I don't need to know the sequences,  
> necessarily, but I would like very much to know what types of sequences  
> are being used, e.g., GPCRs (more specific?), ion channels, etc.

> Diana Hamlet-Cox

> ----- Original Message -----  
> Subject: Toxicology Chip  
> Date: Mon, 19 Jun 2000 18:31:48 -0700  
> From: Diana Hamlet-Cox <dianahc@incyte.com>  
> Organization: Incyte Pharmaceuticals  
> To: grigg@niehs.nih.gov

> Dear Colleague:

> I am doing literature research on the use of expressed genes as  
> pharmacotoxicology markers, and found the Press Release dated February  
> 29, 2000 regarding the work of the NIEHS in this area. I would like to  
> know if there is a resource I can access (or you could provide?) that  
> would give me a list of the 12,000 genes that are on your Human ToxChip  
> Microarray. In particular, I am interested in the criteria used to  
> select sequences for the ToxChip, including any control sequences  
> included in the microarray.

> Thank you for your assistance in this request.

> Diana Hamlet-Cox, Ph.D.  
> Incyte Genomics, Inc.

> --

> =====

> This email message is for the sole use of the intended recipient(s) and  
> may contain confidential and privileged information subject to  
> attorney-client privilege. Any unauthorized review, use, disclosure or  
> distribution is prohibited. If you are not the intended recipient,  
> please contact the sender by reply email and destroy all copies of the  
> original message.

> =====

>  
>  
>

Bengt Bjellqvist\*  
Bodil Basse  
Eyðfinnur Olsen  
Julio E. Celis

Institute of Medical Biochemistry  
and Danish Centre for Human  
Genome Research, Aarhus  
University, Aarhus

## Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [<sup>35</sup>S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human cell types and recorded in the human keratinocyte 2-D gel protein database were identified in the 2-D gel maps and their isoelectric points (pI) were determined using narrow-range IPGs. The latter established a pH scale that allowed comparisons between 2-D gel maps generated either with other IPGs in the first dimension or with different human protein samples. Of the 41 proteins identified, a subset of 18 was defined as suitable to evaluate the correlation between calculated and experimental pI values for polypeptides with known composition. The variance calculated for the discrepancies between calculated and experimental pI values for these proteins was 0.001 pH units. Comparison of the values by the *t*-test for dependent samples (paired test) gave a *p*-level of 0.49, indicating that there is no significant difference between the calculated and experimental pI values. The precision of the calculated values depended on the buffer capacity of the proteins, and on average, it improved with increased buffer capacity. As shown here, the widely available information on protein sequences cannot, *a priori*, be assumed to be sufficient for calculating pI values because post-translational modifications, in particular *N*-terminal blockage, pose a major problem. Of the 36 proteins analyzed in this study, 18–20 were found to be *N*-terminally blocked and of these only 6 were indicated as such in databases. The probability of *N*-terminal blockage depended on the nature of the *N*-terminal group. Twenty six of the proteins had either M, S or A as *N*-terminal amino acids and of these 17–19 were blocked. Only 1 in 10 proteins containing other *N*-terminal groups were blocked.

### 1 Introduction

As compared with carrier ampholyte isoelectric focusing (CA-IEF), the application of immobilized pH gradients (IPGs) in the first dimension in 2-D gel electrophoresis offers improved reproducibility [1] because the nature of the pH gradient makes the resulting focusing positions insensitive to the focusing time [2] and to the type of sample applied [3]. The recently introduced ready-made IPG strips [4] seem to be an ideal substitute for the carrier ampholyte gradients, which until now have been the most commonly used first dimensions in 2-D gel electrophoresis. The availability of standardized first dimensions opens the possibility of comparing 2-D gel maps of various cell types generated in different laboratories, provided that the focusing positions of a number of easily recognizable polypeptide spots common to the cell types

in question are known. Even though this approach is limited to experiments performed with the same standardized IPG, the flexibility provided by IPGs allows the pH gradient to be adjusted to the requirements of a particular experiment.

Exchange and communication of 2-D gel protein data requires a pH scale that is independent of the particular IPG used and by which the results can be described. The introduction of carbamylation trains and the relation of focusing positions to the spots in these trains represented a step forward towards solving the reproducibility problem experienced with carrier ampholyte focusing [5]. Problems associated with the use of carbamylation trains were mainly due to lack of temperature control and to the use of nonequilibrium focusing conditions. Accordingly, the pattern variation involved not only the resulting pH gradients, but also the relative spot positions as related to each other and to spots in the carbamylation trains. Even though the question of reproducibility has, to a large extent, been solved, the carbamylation trains are still not ideal as markers because the spots in the trains do not represent defined entities but rather a large number of differently carbamylated peptides having close pI values. As a result, the spots are large and poorly defined as compared to the ordinary polypeptide spots in 2-D gel maps.

Correspondence: Professor J. E. Celis, Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark.

Abbreviations: CA-IEF, carrier ampholyte-isoelectric focusing; SSP, sample spot number

\* Present address: Pharmacia Biotech AB, S-751 82 Uppsala, Sweden

hardt *et al.* [6] defined the pH gradient in 2-D gel experiments by *pI* markers whose *pI* values were calculated from the amino acid composition. Focusing positions of other polypeptides could be predicted from their position but the *pK* values needed for the *pI* calculations were unknown. Various groups employing this approach do not use the same *pK* values [6, 7] and therefore the *pI* values derived in this way cannot be expected to describe the variation of the hydrogen ion activity. In spite of this fact, it is still possible to make approximate predictions of focusing positions because *pK* values used to define the pH gradient are also used to calculate *pI* values and to predict the focusing positions. Errors in *pK* assignments are therefore compensated. A pH scale which correctly reflects the variation of hydrogen ion activity during focusing should improve the precision of the predictions, but this has never been implemented with CA-IEF focusing as a first dimension in 2-D gel electrophoresis. The main reason for this are problems associated with pH measurements in gels containing high concentrations of urea.

and a focusing temperature of 25°C. We have used commercial nonlinear, wide-range IPG strips which give 2-D gel maps that are closely similar to the ones resulting with the CA-IEF technique used to establish the human keratinocyte database [15]. As an initial step towards interlaboratory comparisons of results obtained with the nonlinear gradient as a first dimension we report here on the focusing positions of 41 known proteins that are common to most human cell types. The pH range covered corresponds to the range in classical CA-IEF 2-D gel electrophoresis and in order to use these proteins as internal standards for comparing 2-D gel maps generated with other IPGs we determined their *pI* values with narrow-range IPGs in the first dimension. We have compared the calculated *versus* experimental *pI* values and show that it is necessary to have further information (absence or presence and nature of posttranslational modifications), in addition to amino acid composition to be able to calculate *pI* values that correspond to the actual experimental values. The *pK* values used for the calculations are provided and the usefulness of *pI* prediction in relation to database information is discussed. Furthermore, we comment on the possibility of using experimentally determined *pI* values to verify the available database information on polypeptide composition.

## 2 Materials and methods

### 2.1 Apparatus and chemicals

Equipment for isoelectric focusing and horizontal SDS electrophoresis (Multiphor<sup>®</sup> II electrophoresis chamber, Immobiline<sup>®</sup> strip tray, Multidrive XL programmable power supply, Macrodrive power supply and Multiemp<sup>®</sup> II) was from Pharmacia LKB Biotechnology AB (Uppsala, Sweden). Vertical second-dimensional gels were run in the home-made equipment described in [15]. The IPG strips with the wide-range nonlinear pH gradient were either Immobiline DryStrip<sup>®</sup> pH 3–10 NL, 180 mm or alternatively 160 mm long IPG strips with a corresponding pH gradient. In both cases the IPG strips were delivered by Pharmacia LKB. Immobiline, Pharmalyte, Ampholine, GelBond as well as PAG film and the ready-made horizontal SDS gels (ExcelGel<sup>®</sup> XL SDS 12–14) were also from Pharmacia LKB. Purified proteins and peptides were from Sigma (St. Louis, MO).

### 2.2 Sample preparation

Preparation and labeling of unfractionated keratinocytes as well as fibroblasts have been described in [16]. Cells were lysed in a solution containing 9.8 M urea, 2% w/v NP-40, 100 mM DTT and 2% v/v Ampholine pH 7–9.

### 2.3 2-D gel electrophoresis

First-dimensional focusing was performed according to Görg *et al.* [2] with some minor modifications, as described in [9]. Rehydration of the IPG strips was made in a solution containing 9.8 M urea, 2% w/v CHAPS, 10 mM DTT and 2% v/v carrier ampholyte mixture. The carrier ampholyte mixture consisted of 2 parts Pharmalyte

can be described from the concentration variation of the immobilized groups, provided that the *pK* values of these groups are known for the conditions prevailing during focusing. To avoid measurements on gels, Gianazza *et al.* [8] suggested the use of *pK* values derived by titration of determined *pK* shifts. Recently, direct determinations of *pK* differences between immobilized spots in IPGs were made by determining *pI*-*pK* values of overlapping narrow-range IPGs [9, 10] and the results confirmed the applicability of the Gianazza approach. A definition of the focusing results in a pH scale, which correctly describes the variation of the hydrogen ion activity for the focusing conditions used, not only allows comparison of 2-D gel maps generated with different *pI* scales, but also opens the possibility for correlating the focusing position of a polypeptide with its composition. Experiments by Bjellqvist *et al.* [9, 10] have implied that pH scales showing good correlation between calculated and experimental *pI* values can be derived for any conditions commonly used for focusing in connection with 2-D gel electrophoresis. These pH scales are defined through the *pK* values of the immobilized spots in the IPG containing gel. To be useful for interlaboratory comparisons, however, the pH scale has to be defined through *pI* values of easily recognizable spots present in the 2-D gel map. So far, *pI* determinations in a useful pH scale, combined with determinations of *pK* values needed for *pI* calculations, have only been made in the pH range 4.5–6.5 at 10°C [9]. CA-IEF focusing as described by O'Farrell [11] does not control the temperature of the first dimension, which can be expected to be slightly above room temperature. With IPGs, the temperature commonly used is about 20°C [4, 12] or 25°C [13] and this is a critical parameter that needs to be controlled [14].

The present work was designed to compare 2-D gel maps of different cell types in a laboratory applying both CA-IEF and IPG focusing at a common temperature. To this end we have generated 2-D gel maps of proteins from noncultured, unfractionated normal human epidermal keratinocytes with IPG in the first dimension

-6.5. 1 part Ampholine pH 6-8 and 1 part Pharmalyte H 8-10.5. Usually, cathodic sample application was used and the samples were diluted 2-20 times in a solution containing 9.8 M urea, 4% w/v CHAPS, 1% w/v DTT and 35 mM Tris base. For acidic application, the Tris-base was substituted with 100 mM acetic acid. The degree of dilution and sample volume (20-100  $\mu$ L) depended on the particular sample and the IPG, and whether visualization of the proteins was to be done by Coomassie Brilliant Blue or silver staining. With the wide-range non-linear IPG, 10-30  $\mu$ g of total protein was loaded for silver staining and 100-200  $\mu$ g for Coomassie staining. Focusing was done overnight with voltages in the range of 45-60 kVh with 160 mm long strips and 50-70 kVh with 180 mm long strips. Solubilization of polypeptides and blocking of -SH groups prior to the second-dimensional run, as well as loading on the second-dimensional gel was done as described in [9]. The stacking gel was omitted and 5-10 mm were left at the top of the second-dimensional gel for applying the IPG strip. The space was filled with electrode buffer containing 0.5% w/v agarose. Casting, running, staining and autoradiography were carried out as described in [15].

#### 4 Experimental determination of pI values

The determination of the pI differences between Immobilines pI 4.6, pI 6.2 and pI 7.0 necessary for the calibration of the pH scale at 25°C in 9.8 M urea was done as described in [9] with the same narrow-range IPGs. The pH scale was defined by setting the pI value of Immobililine pI 4.6 equal to 4.61 [9] and the determined differences gave the pI values of Immobilines pI 6.2 and pI 7.0, equal to 5.73 and 6.54, respectively. The pI differences found are in good agreement with values determined from [17] and [8] by extrapolation to 9.8 M urea concentration. As in [9], additional narrow-range recipes have been used for determining pI values. With narrow-range IPGs extending to pH values higher than the pI value of Immobililine pI 7.0, anodic sample application is used with acetic acid added to the sample solution. Otherwise, cathodic sample application was used with the same sample buffer as for wide-range IPGs.

#### 5 Protein compositions used for pI calculations

With the exception of vimentin, protein compositions were taken from the Swiss-Prot database [18]. For vimentin, we used the data from [19], where the amino acid at position 41 is a D instead of a S. Information in the Swiss-Prot database on phosphorylation has been disregarded because it was known from earlier studies (J. E. Celis, unpublished results) that the spots in question corresponded to the unphosphorylated forms of the peptides.

#### 6 Calculation of pI values

For the pI calculations it was assumed that the same pI value could be used for an amino acid residue in all peptides and in all positions in the peptide except N- or C-terminally placed amino acids. For the pI values of the N-terminal amino groups the effect of the

different substituents on the  $\alpha$ -carbon were taken into account. The calculations of pI values were made with the aid of the IPG-maker program [20].

#### 2.7 pK values used for pI calculations

For the carboxyl terminal group and internal glutamyl and aspartyl residues the same pK values were used as in [9]. For C-terminal glutamyl and aspartyl residues, separate pK values were derived with the aid of the Taft equations [9, 21]. The pK values of histidyl groups were calculated from the pI values of human carbonic anhydrase I as in [9]. For N-terminal glycine a pK value of 7.50 was used. The pK shift caused by a substituent on the  $\alpha$ -carbon was assumed to be identical with the pK shift the substituent caused for the amino group in the amino acid, i.e. 2.28 pH units were subtracted from the pK values for the amino groups in the amino acids given in [22, 23]. The approximate pK value of 9 for the cystenyl group was taken from [24]. For tyrosyl and arginyl groups we used the pK values for the amino acids [22, 23]. For lysyl groups the effect of high urea concentration on amino groups was taken into account and 0.5 pH units were subtracted from the amino acid pK value. These last three pK values are far from the pH range under study and the results found would have been the same if lysyl and arginyl groups were assumed to be fully ionized while the ionization of tyrosyl groups were neglected. A complete list of the pK values used is given in Table 1.

Table 1. pK Values used for the ionizable groups in peptides  
9.8 M urea, 25°C

Ionizable group	pK
C-terminal	3.55
N-terminal	
Ala	7.50
Met	7.00
Ser	6.93
Pro	8.36
Thr	6.82
Val	7.34
Glu	7.70
Internal	
Asp	4.05
Glu	4.45
His	5.98
Cys	9
Tyr	10
Lys	10
Arg	12
C-terminal side chain groups	
Asp	4.55
Glu	4.75

#### 2.8 Statistical analysis

Statistical comparisons of the experimental and calculated pI values were done on an Apple Macintosh IIsi using the statistical package Statistica/Mac, release 3.0b (from StatSoft Inc., Tulsa, Oklahoma). Calculated and experimental pI values were compared by the t-test for

correlated samples (paired *t*-test). The normality of *pI* differences was estimated graphically by probability plots. The variances of the data presented here and the similar data on plasma and liver proteins in [9] were compared by the *F*-test.

### 3 Results and discussion

#### 3.1 Identification of polypeptides and *pI* determinations

The 2-D gel maps of [<sup>35</sup>S]methionine-labeled proteins from noncultured, unfractionated normal human kerati-

nocytes, focused with the nonlinear, wide-range IPG and CA-IEF pH gradients in the first dimension, are shown in Figs. 1 and 2, respectively. The IPG extends to higher *pH* values but otherwise the two patterns are very similar and most of the spots in the IPG pattern can be directly related to the corresponding spots in the CA-IEF gel. To obtain comparable patterns it was important to keep the focusing temperature as similar as possible. Compared to other studies [1-4, 9, 10, 12-14], we increased the urea concentration in the focusing gel to 9.8 M because keratins streaked badly in the focusing dimension when 8 M urea was used, presumably due to

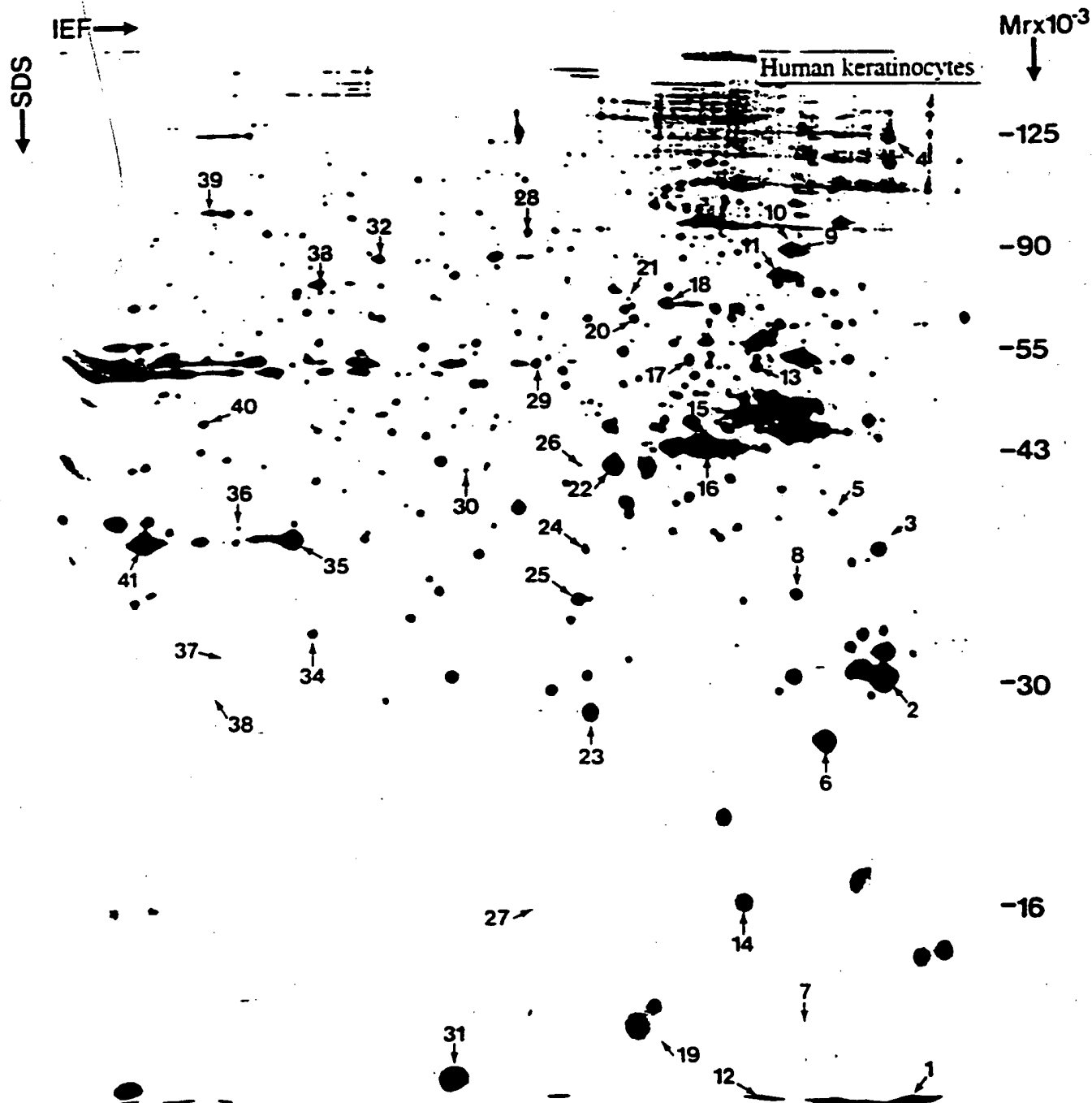


Figure 1. 2-D gel protein map of [<sup>35</sup>S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

aggregates of acidic and basic keratins. An increase in urea concentration to 9 M or more eliminated these streaks; apart from this effect, no other major changes in the focusing positions were observed. In Fig. 1 we have indicated the positions of 41 known proteins from the human keratinocyte 2-D gel database that are most likely common to most human cell types. The choice was made because these proteins are easy to identify with certainty. With the exception of stratifin (spot 2), involucrin (spot 4) and keratin 14 (spot 15), which are all

epithelial markers, these proteins are also present in human fibroblasts (Fig. 3) and lymphocytes (results not shown), and therefore can be used as landmarks for comparing 2-D gel maps derived from different cell types. In Table 2 the 41 proteins are listed together with their sample spot numbers (SSP) in the human keratinocyte protein database and pI values determined in 2-D gel maps generated with narrow-range IPGs in the first dimension.

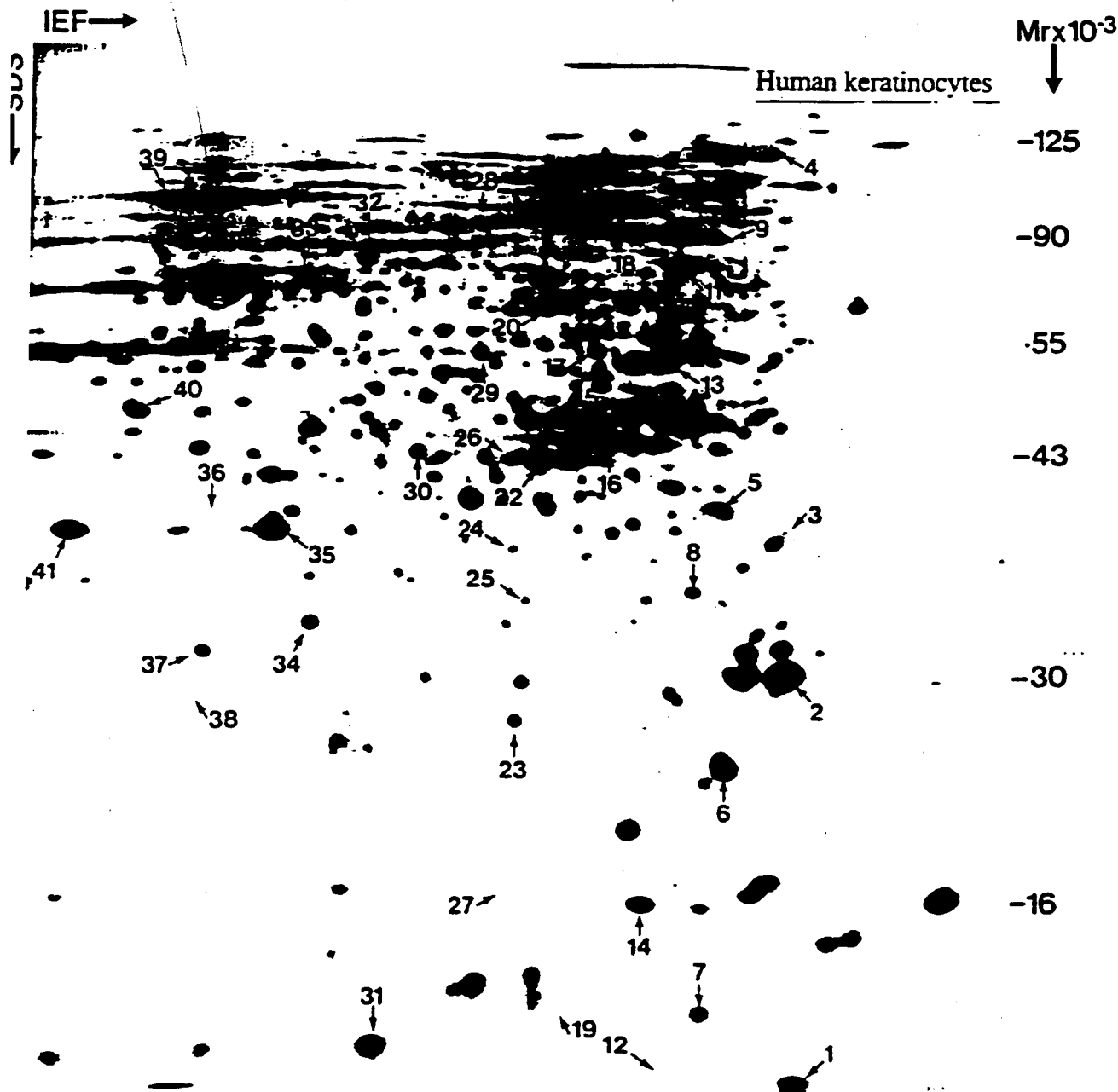


Figure 2. 2-D gel protein map of [ $^{35}$ S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with IEF in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

Table 2. Proteins from the human keratinocyte database localized in 2-D gels run with IPGs as first dimension

Number in Protein name Figs. 1-3	Protein name	H-1 SSP number <sup>a)</sup>	Experimental pI value	Calculated pI value	Discrepancy (pI units)	Calculated net charge at experimental pI value	Bulter capacity charge units per pI unit	N-terminal blockage	Recalculated for suspected blockage	Discrepancy pI units	Net charge	Swiss-Prot accession number
1	CaN 19	9027	4.46									
2	Stratelin, bovine 14.3.3 related protein	9109	4.58									P12004
3	Proliferating nuclear antigen (PCNA)/cyclin	9226	4.58									P07476
4	Involucrin	9703	4.63	4.57	-0.01	-0.1	20.8					P06748
5	Nucleolar protein B23	8207	4.75	4.63	-0.12	-0.3	70.1					P13693
6	Transcriptionally controlled tumor protein	8114	4.79	4.84	0.05	0.6	13.1					P10599
7	Thioredoxin	8006	4.86	4.82	-0.04	-0.3	7.1					P08758
8	Annexin V	8213	4.89	4.88	-0.01	-0.1	20.3					P07900
9	Heat shock protein 90- $\beta$	8611	4.95	4.94	-0.01	-0.5	56.2					P08238
10	Heat shock protein 90- $\alpha$	2629	4.97	4.97	0.00	0.2	53.6					P11021
11	Glucose regulated protein 78 (GRP)	8515	4.99	4.98	-0.01	-0.6	37.5					P06703
12	Calycalin	8017	5.02	5.02	0.00	1.3	3.6					P08670
13	Vimentin	8417	5.05	5.06	0.01	0.2	27.1		5.09	0.07	0.3	P10159
14	Initiation factor 4D	8016	5.05	5.08	0.03	0.2	7.6					P02533
15	Keratin 14	7305	5.08	5.09	0.01	0.2	21.0					P02570
16	$\beta$ -Actin	7316	5.21	5.21	0.00	0.06	13.3					P10809
17	Heat shock protein 60	6403	5.23	5.24	0.01	0.1	17.5		5.32	0.04	0.8	P11142
18	Heat shock cognate 71kD	6501	5.28	5.37	0.09	1.8	18.1					P01040
19	Cystatin	6011	5.30	5.38	0.08	0.2	3.0		5.36	0.02	0.3	P13797
20	I-plasin	6112	5.34	5.31	-0.03	1.3	23.3					P08433
21	Callectrin	5628	5.35	5.37	0.02	0.5	10.7		5.37	0.01	0.07	P05120
22	Plasminogen activator inhibitor-2	6314	5.38	5.46	0.08	0.9	3.9					P09211
23	Glutathione S-transferase $\alpha$	5101	5.43	5.44	0.01	0.08	8.7					P13928
24	Annexin VIII	5213	5.45	5.36	-0.09	1.4	8.4		5.46	0.01	0.05	P12429
25	Annexin III	5204	5.46	5.63	0.17	1.8	10.8		5.52	0.06	0.5	P08613
26	Adenosine deaminase	5305	5.47	5.63	0.16	1.8	6.6		5.54	0.07	0.8	P16949
27	Stathmin	5001	5.55	5.61	0.06	0.4	16.5					P06396
28	Gelsolin, cytoplasmic	5608	5.59	5.58	-0.01	-0.1						P15111
29	Ra phosphoinositide specific protein homolog	5410	5.62									P26018
30	Flasase inhibitor	4314	5.74									P00191
31	S100, calgrizin	4006	5.75	5.95	-0.04	-0.5	13.2					P01081
32	Cytellin, c/rrn	3504	5.99 <sup>1</sup>	6.09	-0.02	-0.2	9.8		6.28	0.17	0.9	P15171
33	Moexin	3515	6.11	6.45	0.34	1.8	4.4		6.11	0.15	0.6	P08669
34	Purine nucleoside phosphorylase	2108	6.11	6.45	0.34	1.6	2.5		6.16	0.01	0.2	P11619
35	Annexin I	2216	6.18	6.64 <sup>2</sup>	0.46	0.7	4.2		6.16	0.00	0.0	P08669
36	Aldose reductase	1202	6.40	6.55	0.15	0.9	2.6					P11619
37	Phosphoglycerate mutase (B form)	1107	6.46	6.75	0.29	-0.04	2.3					P06711
38	Triphosphatase isomerase	1111	6.53	6.51	-0.02	-0.5	9.8					P06711
39	Elongation factor 2	1610	6.43	6.38	-0.05	1.0	2.2		6.75	0.11	0.1	P06711
40	$\alpha$ -Enolase	1325	6.62	6.99	0.37	1.0	0.9					P01135
41	Annexin II	210	7.30	7.36	0.06	0.05						

a) SSP number in the keratinocyte database [15]

b) Peptides N-terminally sequenced as liver proteins [3]

c) Peptides given as N-terminally blocked in Swiss-Prot database



### 3.2 Comparison between the determined and calculated pI values for human keratinocyte proteins

Thirty six of the 41 proteins listed in Table 2 are found in the Swiss-Prot database. Contrary to the plasma and liver proteins used in [9], the pI calculations on the proteins used in this study posed some problems that reflected the way in which they were characterized. The

proteins used by Bjellqvist *et al.* [9] were either very abundant and well-characterized plasma proteins or they were identified by N-terminal sequencing and, therefore, the nature of the N-terminals (acetylated or non-acetylated) was in both cases known. The proteins used in this study have all been characterized by internal sequencing [7] and it is known that N-terminal acetylation occurs with high frequency in eukaryotes.

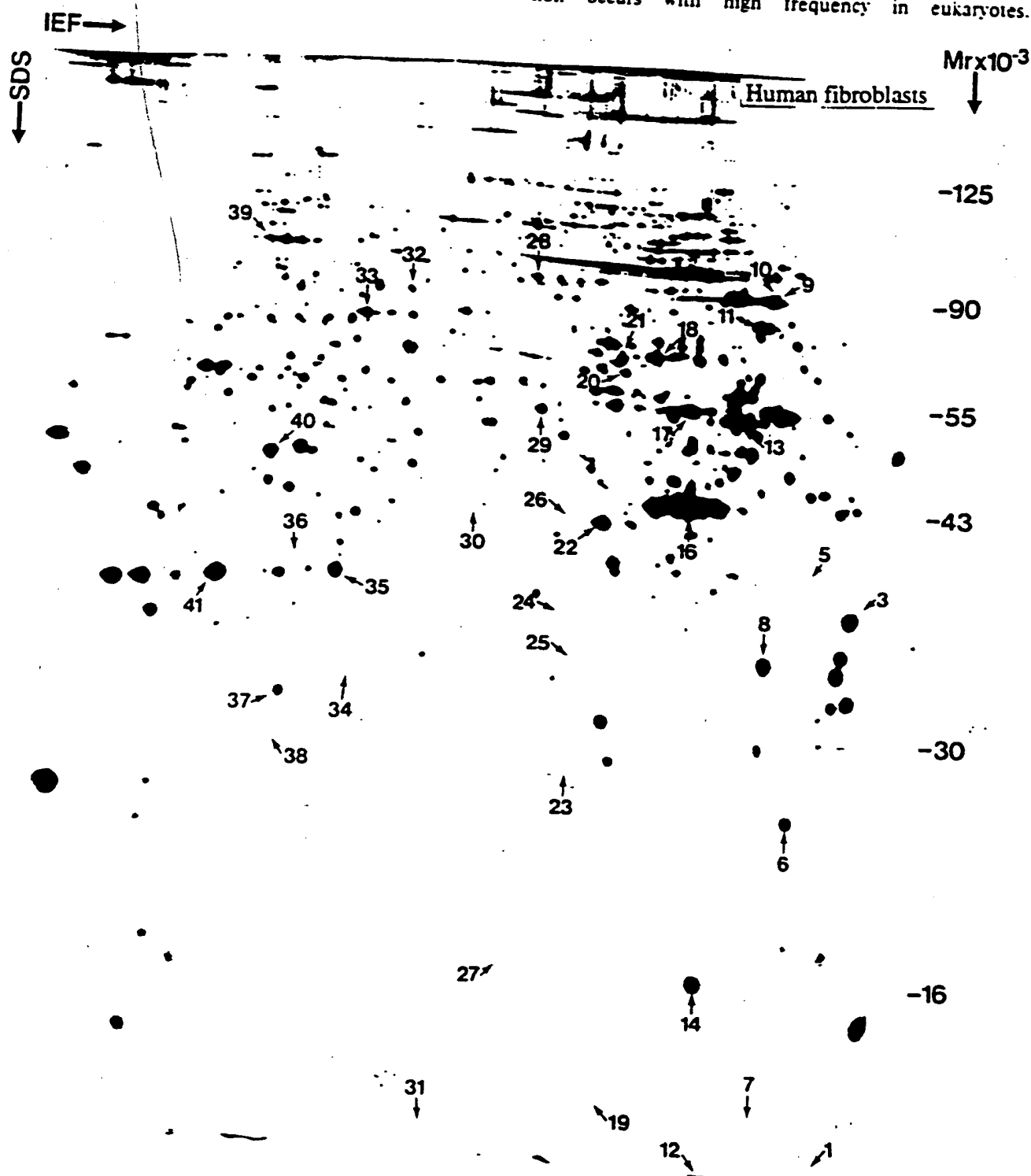


Figure 3. 2-D protein map of [35S]methionine-labeled proteins from normal human fibroblasts focused with the nonlinear, wide-range IPG in the 1st dimension. The position of the 41 proteins analyzed in this study is indicated.

According to Brown and Robert [25], proteins with acetylated *N*-terminals correspond in weight to approximately 80% of the soluble protein in ascites cells. Based on results from *N*-terminal sequencing, at least 40% of the spots in the human liver protein 2-D gel map appear to be blocked [3]. The corresponding number, derived from 107 spots in the 2-D gel map of human T-lymphocyte proteins, falls between 60 and 65% (J. Strahler, personal communication). Information concerning *N*-terminal blockage is not normally available, and in the Swiss-Prot database only 6 of the 36 keratinocyte proteins are specified as *N*-terminally blocked. We have, within the present material, defined 18 proteins for which the *N*-terminals are very likely to be correctly described. Six of these proteins are listed in the Swiss-Prot database as *N*-terminally blocked, four represent proteins which appear in the human liver 2-D gel map and have been *N*-terminally sequenced as liver proteins [3] and the remaining eight have *N*-terminal groups other than M, S and A, i.e. *N*-terminals for which *N*-acetylation is uncommon [26]. In Figs. 4A, B, C and D *pI* values calculated from Swiss Prot database information are plotted against the experi-

mentally determined *pI* values for all the keratinocyte proteins listed in Table 2 and for the 18 selected proteins, as well as for the plasma and liver proteins (data from [9] valid for 10°C)\*.

The calculations show that without knowledge of the status of the *N*-terminal group, precise predictions of *pI* values for eukaryotic proteins cannot be achieved based on the information available in Swiss-Prot and similar databases. However, for proteins where the *N*-terminal status is known, we find good correlation between predicted and experimental *pI* values. When the variance of the *pI* discrepancies and the variance of calculated charges at the experimental *pI* values derived from the present data set are compared with the corresponding

\* There are four plots: (A) the 36 polypeptides from normal human keratinocytes (no corrections), (B) the 36 polypeptides from Fig. 4A where *pI* values have been recalculated for 12 polypeptides with M, S and A as *N*-terminally assumed blocked, based on calculated charge, (C) the 18 selected polypeptides with information on the *N*-terminal configuration, and (D) plasma and liver proteins.

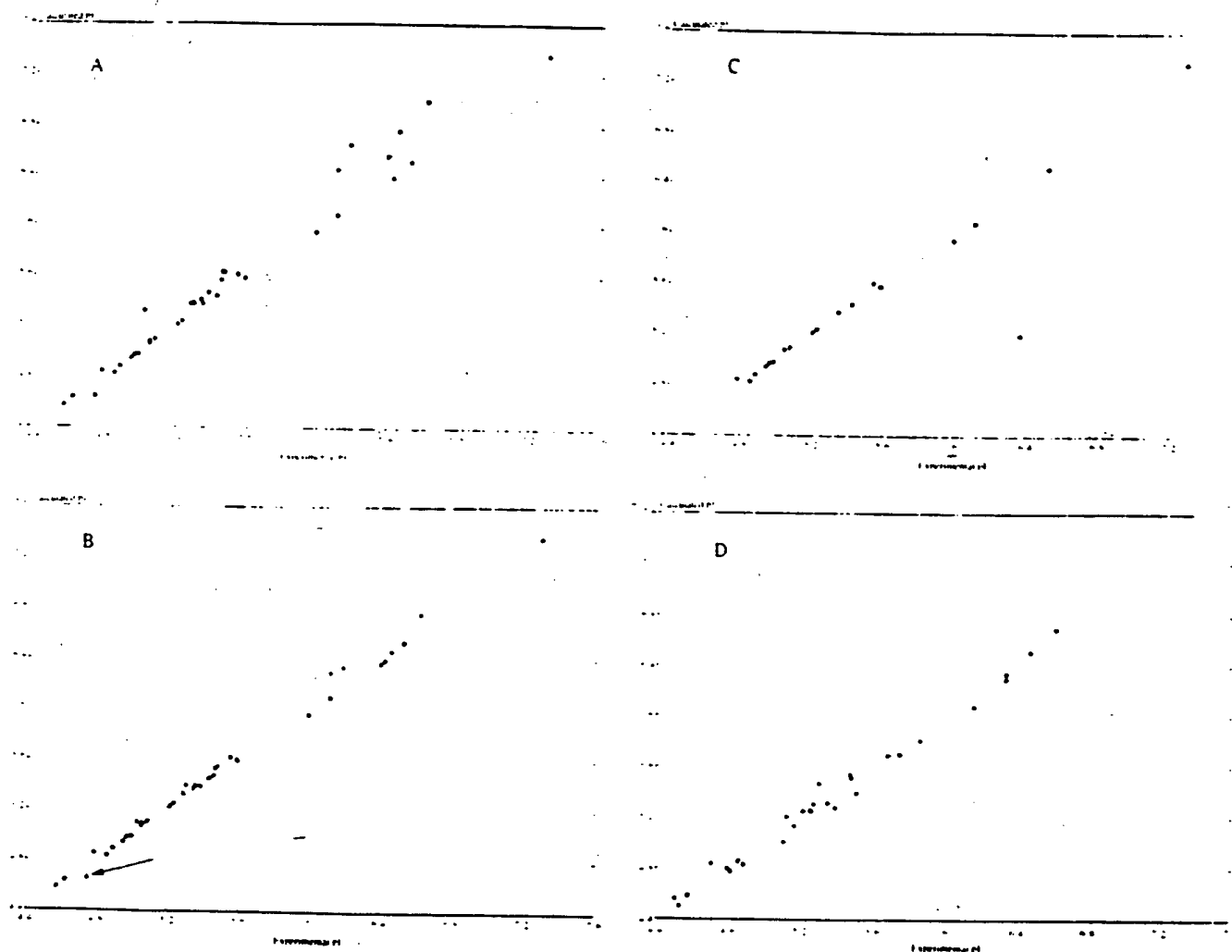


Figure 4. Calculated vs. experimental *pI* values. Lines are fitted using the least squares criterion. (A) 36 polypeptides from normal human keratinocytes (no corrections). (B) 36 polypeptides from Fig. 4A (including the 18 marker polypeptides) where *pI* values have been recalculated assuming *N*-terminal blockage; x indicates recalculated *pI* values; nucleolar protein B23 is indicated with an arrow. (C) 18 polypeptides with information on *N*-terminal configuration and (D) plasma and liver proteins.

ies derived from the data on plasma and liver proteins in [9] (Table 3), the present data are found to result in larger variances for the values of both  $pI$  discrepancies calculated charge at the experimental  $pI$  value when information on posttranslational modification is taken into consideration. Correction for possible N-acetylation of 12 polypeptides with M, S and A as N-terminal results in a smaller variance of  $pI$  discrepancies, although not significantly different from values derived in [9], whereas the variance of the calculated charge at experimental  $pI$  value is significantly higher. For the selected proteins the variance for the  $pI$  discrepancies is significantly smaller than for the data in [9]; however, the corresponding value for calculated charge at the experimental  $pI$  value does not improve to the same extent. This, we believe, reflects another difference between the two sets of proteins used for the calculations. Based on spot distributions in 2-D gel maps, the set of proteins used here has a molecular weight distribution that is more representative of the patterns observed in mammalian cells. In the study by Bjellqvist [9] most of the high molecular weight plasma proteins had to be excluded due to their unknown content of sialic acid which made the proteins analyzed in this study heavily biased towards low molecular weight proteins. The buffer capacity of proteins normally increases with the protein's molecular weight, and the average buffer capacity of the presently selected proteins with known N-terminals is 18 charge units/pH unit, compared to the corresponding value for the proteins used in [9] of only 9 charge units/pH unit. High buffer capacity is expected to improve the agreement between calculated and experimental  $pI$  values. Inspection of the data presented in Table 2 for the polypeptides with known N-terminals verifies the importance of buffer capacity. For 8 polypeptides having buffer capacities higher than 15 charge units/pH unit, the calculations in all cases yielded  $pI$  discrepancies with absolute values of less than 0.02 pH units. The largest discrepancy, 0.06 pH units, was observed for annexin II and min. proteins which have low buffer capacity: 0.9

and 6.6 charge units/pH unit, respectively. The probability that the focusing position of a protein with known composition will fall within a certain distance from the calculated  $pI$  value therefore cannot be predicted by the variance alone. The buffer capacity of the specific protein must be taken into consideration as well. As indicated by the decrease of the variance of calculated charges at the experimental  $pI$  value for the selected proteins, the observed improvement can not solely be due to the higher buffer capacity of the keratinocyte proteins. The two studies relate to different experimental conditions. Good agreement between experimental and calculated  $pI$  values implies that the proteins are defolded and a factor that may contribute to the observed improvement is a more complete defolding of proteins caused by the higher temperature and urea concentration used in this study.

The data indicated that the precision with which  $pI$  values can be predicted for polypeptides with high buffer capacity is better than the precision with which experimental  $pI$  values can be determined. If the pH is defined through the  $pK$  values of the immobilized groups in the IPG containing gel, the precision of the experimentally calculated data will depend on the pH difference between the  $pI$  and the  $pK$  value of the immobilized group with the closest  $pK$ . For the present study this will give  $pI$  determinations with a precision varying in the range of  $\pm 0.02$ – $0.05$  pH units [9]. The good agreement observed between the calculated and experimental  $pI$  values is due to the fact that errors are mainly systematic and, as discussed in [9], they will largely be cancelled out in the calculations. A pH scale defined through the presently determined  $pI$  values will not necessarily reflect the variation of the hydrogen ion activity during the focusing step in an optimal way, but it still allows precise predictions of focusing positions for polypeptides with known compositions, including information on posttranslational modifications. Calculated net charge at the experimentally found isoelectric point defined in this scale will serve as a tool to verify that the polypeptide

3. Mean values and variances for the difference (experimental  $pI$ -calculated  $pI$ ) in pH units and calculated charges at the experimental  $pI$  values, respectively

Set of proteins	Plasma and liver proteins (18 M urea, 10°C)				Keratinocyte proteins (9.8 M urea, 25°C)			
	All peptides		All peptides after correction for N-acetylation		Known N-terminal configuration (or very likely configuration)			
	29	36	36	18				
Experimental $pI$ -calculated $pI$	Mean -0.011	Variance 0.005	Mean 0.072	Variance 0.017	Mean 0.019	Variance 0.003	Mean 0.005	Variance 0.001
Calculated $pI$ discrepancy	1	3.4	1.67	5				
Calculated charge at the experimental $pI$ value	Mean -0.070	Variance 0.227	Mean 0.321	Variance 0.871	Mean 0.009	Variance 0.444	Mean -0.014	Variance 0.109
Calculated charge at the experimental $pI$ value	1	3.8	1.96	2.08				
Calculated charge at the experimental $pI$ value	Mean 0.5	Variance 0.0002	Mean 0.0338	Variance 0.0536				

Comparison to the data in [9].  $F = S_1^2/S_2^2$ , where  $S_1^2$  is the larger of the two variances ( $F_{v_1, v_2} \geq F$ -value), where  $v_1$  and  $v_2$  are the degrees of freedom for  $s_1$  and  $s_2$ , respectively

composition used in the calculation is correct and complete. Exceptions to this are proteins such as involucrin and heat shock protein 90 that have very high buffer capacities. Introduction of an extra charge unit into these proteins will only result in  $pI$  shifts falling in the range of 0.01–0.02 pH units and the effect is that the quality of the  $pH$  definition – the precision by which  $pK$  values used in the calculations are given and the precision of experimental  $pI$  values in these cases – will limit the possibilities to verify polypeptide composition based on the experimental  $pI$  value.

Statistical comparison of experimental and calculated  $pI$  values was done using the  $t$ -test for dependent samples and normality of the discrepancies was estimated by probability plots. For the 36 proteins, the  $p$ -level is 0.0021, indicating that a result like this is unlikely to be a chance effect and must be assumed to represent a real difference. After correction for the most likely  $N$ -terminal configuration, the  $p$ -level is 0.043 and cannot be accepted as representing the same population since the  $p$ -level is less than 0.05 – the traditional  $p$ -limit of statistical significance. For the 18 proteins with a known or very likely  $N$ -terminal configuration the  $t$ -test gave a  $p$ -level of 0.49, which verifies that the experimental and calculated  $pI$  values are not significantly different.

Besides showing that  $pI$  values for denatured proteins with known compositions can be calculated with a high degree of precision from average  $pK$  values, the results also provide strong support for the notion that  $N$ -terminal blockage heavily depends on the nature of the  $N$ -terminal groups [26]. The results seem to indicate that with  $N$ -terminals other than M, S and A, only a few proteins have blocked  $N$ -terminals (1 out of 10 proteins in the present study), while it can be inferred from the data presented in Table 2 that a majority of the proteins with M, S and A as  $N$ -terminal are blocked. After correction for the effect of suspected  $N$ -terminal blockage there is only one protein (nucleolar protein B23) out of the 36 used in this study, which, in spite of a high buffer capacity, has a marked difference of 0.11 pH units between predicted and determined  $pI$  values (Fig. 4B); this corresponds to 3 charge units due to the high buffer capacity of this protein. This discrepancy in  $pI$  prediction and calculation of net charge at the  $pI$  is probably not due to deficiencies in the database information but instead reflects a shortcoming of the model used for  $pI$  calculations. Nucleolar protein B23 contains a domain extremely rich in aspartic and glutamic acid residues (Table 4), in which 26 out of 28 amino acid residues from position 161 to 188 are either a D or an E. A calculation based on the use of average  $pK$  values uninfluenced by the charged neighboring amino acid residues cannot be expected to correctly describe the  $pI$  value with almost half of the acidic groups packed

together into a highly negatively charged region. This limitation caused by calculations based on average  $pK$  values does not severely limit the usefulness of the approach since a search through Swiss-Prot shows that this type of D/E-rich motif is uncommon, and the existence of a highly charged region is immediately apparent upon inspection of the amino acid sequence.

The quality of the information available in databases, especially concerning posttranslational modifications, is a major problem when the data is to be used for  $pI$  predictions. The  $p$ -level of 0.043 found for all 36 proteins after correction for  $N$ -acetylation, shows that this problem is not only limited to  $N$ -terminal blockage and the very good agreement found for the eighteen polypeptides, with assumingly correctly described  $N$ -terminal (Fig. 4C), must be regarded as an exception from this point of view.  $N$ -Terminal blockage is generally the main problem in relation to  $pI$  predictions for eukaryotic proteins. Of the 36 keratinocyte proteins analyzed, 18–20 are suspected to be  $N$ -terminally blocked (6 proteins blocked according to Swiss-Prot, 12 proteins with M, S or A as  $N$ -terminal and assumingly blocked based on the calculated charge, and two proteins, involucrin and nucleolar protein B23, with M as  $N$ -terminal for which the data does not allow any conclusion). This is in reasonable agreement with the conclusions based on the  $N$ -terminal sequencing data derived in connection with 2-D gel electrophoresis.  $N$ -terminal blockage can be suspected for 17–19 of the 26 proteins with M, S or A as  $N$ -terminal, while only 1 in 10 proteins with other  $N$ -terminal groups are blocked. The information that the frequency of  $N$ -terminal blockage is strongly related to the nature of the  $N$ -terminal group will be of some help in connection with  $pI$  predictions based on database information. However, without information from other sources, an uncertainty will always remain as to whether the  $N$ -terminal charge should be included in the  $pI$  calculation.

#### 4 Concluding remarks

The data presented here lays the foundation for comparing 2-D gel protein maps of different cell types generated with nonlinear, wide-range IPGs in the first dimension. The focusing positions of 41 polypeptides common to most human cell types have been described in a  $pH$  scale that allows focusing positions to be predicted with a high degree of accuracy, provided that the composition of the polypeptides are known and that information on posttranslational modifications are available. For polypeptides with a very high buffer capacity, the limiting factor is the precision with which experimental  $pH$  values can be determined rather than the precision of the calculations. Possible deficiencies in the  $pH$  scale description of the variation of the hydrogen ion activity has, at least at the present state, no consequences for its practical use. The major limitation in connection with predictions of focusing positions from polypeptide compositions is the quality of existing data on protein compositions, especially concerning posttranslational modifications. Amino acid sequences have been reasonably easy to obtain, while posttranslational modifications

Table 4. Amino acid sequence of nucleolar phosphoprotein B23

1	MEDSDQDS	FAFQDQDS	DEFAQDQD	FAFQDQDS	DEFAQDQD
51	AGNCEQNT	ENQCEQNT	ENQCEQNT	ENQCEQNT	ENQCEQNT
101	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS
151	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS
201	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS
251	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS	FAFQDQDS

have been difficult and work-intensive to determine. Recent developments in the field of mass spectrometry are fast changing this situation and within the next years we can expect a surge in reliable data in this area. While waiting this development, verification of correctness and completeness of available information on polypeptide composition can be provided by experimental  $pI$  values in a  $pH$  scale based on the  $pI$  values determined in this study. So far, our data cover the  $pH$  range below  $pH \approx 7.5$ . The basic  $pH$  range covered by NEPHGE as first dimension will be covered in forthcoming work.

Received December 29, 1993

## References

- [1] Gianazza, E., Astrua-Testori, S., Caccia, P., Giaccon, P., Quaglia, L., Righetti, P. G., *Electrophoresis* 1986, 7, 76-83.
- [2] Gorg, A., Postel, W., Guntner, S., *Electrophoresis* 1988, 9, 531-546.
- [3] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, 13, 992-1001.
- [4] *Immobiline Dr-Strip Kit for 2-D Electrophoresis: Instructions*, Pharmacia LKB Biotechnology AB, Uppsala 1993.
- [5] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, 93, 312-320.
- [6] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E., Phillips, T. A., *Electrophoresis* 1989, 10, 116-121.
- [7] Rasmussen, H. H., Damme, J. V., Puype, M., Gesser, B., Celis, J. E., Vandekerckhove, J., *Electrophoresis* 1992, 13, 960-969.
- [8] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, 4, 321-326.
- [9] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1023-1031.
- [10] Bjellqvist, B., Pasquali, C., Ravier, C., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1357-1365.
- [11] O'Farrell, P. H., *J. Biol. Chem.* 1975, 250, 4007-4021.
- [12] Gorg, A., *Biochem. Soc. Transactions* 1993, 21, 136-137.
- [13] Hanash, S. M., Strahler, J. R., Neel, J. V., Hailat, N., Mainem, R., Keim, D., Zhu, X. X., Wagner, D., Gage, D. A., Watson, J. T., *Proc. Natl. Acad. Sci. USA* 1991, 88, 5709-5713.
- [14] Gorg, A., Postel, W., Friedrich, C., Kuick, R., Strahler, J. R., Hanash, S. M., *Electrophoresis* 1991, 12, e53-e58.
- [15] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honore, B., Dejgaard, K., Gromov, P., Hoffmann, H. J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kiergaard, I., Puype, M., Van Damme, J., Delay, B., Vandekerckhove, J., *Electrophoresis* 1993, 14, 1091-1198.
- [16] Celis, J. E., Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., Celis, A., Lauridsen, J. B., Basse, B., Ratz, G. P., Andersen, A., Walbum, E., Brandstrup, B., Pedersen, P. S., Brandt, N. J., Puype, M., Van Damme, J., Vandekerckhove, J., *Electrophoresis* 1991, 12, 802-872.
- [17] Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Gorg, A., Postel, W., Westermeier, R., *J. Biochem. Biophys. Methods* 1982, 6, 317-333.
- [18] Bairoch, A., Boeckman, B., *Nucleic Acids Res.* 1991, 19, 2247-2249.
- [19] Honore, B., Madsen, P., Basse, B., Andersen, A., Walbum, E., Celis, J. E., Leffers, H., *Nucleic Acids Res.* 1990, 18, 6692.
- [20] Altland, K., *Electrophoresis* 1990, 11, 140-147.
- [21] Perrin, D. D., Dempsey, B., Serjani, E. P., *pKa Predictions for Organic Acids and Bases*, Chapman and Hall Ltd., London 1981.
- [22] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Butterworths, London 1965.
- [23] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Supplement 1972, Butterworths, London 1972.
- [24] Altland, K., Becher, P., Rossman, U., Bjellqvist, B., *Electrophoresis* 1988, 9, 474-485.
- [25] Brown, J. L., Robert, W. K., *J. Biol. Chem.* 1976, 251, 1009-1014.
- [26] Persson, B., Flinta, C., Heine, G., Jörnvall, H., *Eur. J. Biochem.* 1985, 152, 523-527.

# GENES

---

**Benjamin Lewin**

Oxford New York Tokyo  
Oxford University Press  
1997

*Oxford University Press, Great Clarendon Street, Oxford OX2 6DP  
Oxford New York  
Athens Auckland Bangkok Bogota Bombay Buenos Aires  
Calcutta Cape Town Dar es Salaam Delhi Florence Hong Kong  
Istanbul Karachi Kuala Lumpur Madras Madrid Melbourne  
Mexico City Nairobi Paris Singapore Taipei Tokyo Toronto  
and associated companies in  
Berlin Ibadan*

*Oxford is a trade mark of Oxford University Press*

*Published in the United States  
by Oxford University Press, Inc., New York*

*© Oxford University Press and Cell Press, 1997*

*All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press. Within the UK, exceptions are allowed in respect of any fair dealing for the purpose of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms and in other countries should be sent to the Rights Department, Oxford University Press, at the address above.*

*This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, re-sold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.*

*A catalogue record for this book is available from the British Library*

*Library of Congress Cataloging in Publication Data  
(Data available)*

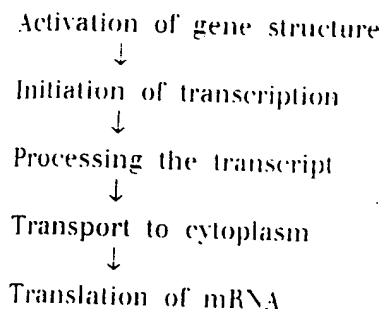
*Typeset by Wytvern Typesetting Ltd, Bristol*

*Printed in The United States of America*

# CHAPTER 29

## Regulation of transcription

The phenotypic differences that distinguish the various kinds of cells in a higher eukaryote are largely due to differences in the expression of genes that code for proteins, that is, those transcribed by RNA polymerase II. In principle, the expression of these genes might be regulated at any one of several stages. The concept of the "level of control" implies that gene expression is not necessarily an automatic process once it has begun. It could be regulated in a gene-specific way at any one of several sequential steps. We can distinguish (at least) five potential control points, forming the series:



The existence of the first step is implied by the discovery that genes may exist in either of two structural conditions. Relative to the state of most of the genome, genes are found in an "active" state in the cells in which they are expressed (see Chapter 27). The change of structure is distinct from the act of transcription, and indicates that the gene is "transcribable." This suggests that acquisition of the "active" structure must be the first step in gene expression.

Transcription of a gene in the active state is

controlled at the stage of initiation, that is, by the interaction of RNA polymerase with its promoter. This is now becoming susceptible to analysis in the *in vitro* systems (see Chapter 28). For most genes, this is a major control point; probably it is the most common level of regulation.

There is at present no evidence for control at subsequent stages of transcription in eukaryotic cells, for example, via antitermination mechanisms.

The primary transcript is modified by capping at the 5' end, and usually also by polyadenylation at the 3' end. Introns must be spliced out from the transcripts of interrupted genes. The mature RNA must be exported from the nucleus to the cytoplasm. Regulation of gene expression by selection of sequences at the level of nuclear RNA might involve any or all of these stages, but the one for which we have most evidence concerns changes in splicing: some genes are expressed by means of alternative splicing patterns whose regulation controls the type of protein product (see Chapter 30).

Finally, the translation of an mRNA in the cytoplasm can be specifically controlled. There is little evidence for the employment of this mechanism in adult somatic cells, but it does occur in some embryonic situations, as described in Chapter 7. The mechanism is presumed to involve the blocking of initiation of translation of some mRNAs by specific protein factors.

But having acknowledged that control of gene expression can occur at multiple stages, and that production of RNA cannot inevitably be equated with production of protein, it is clear



that the overwhelming majority of regulatory events occur at the initiation of transcription. Regulation of tissue-specific gene transcription lies at the heart of eukaryotic differentiation; indeed, we see examples in Chapter 38 in which proteins that regulate embryonic development prove to be transcription factors. A regulatory transcription factor serves to provide

common control of a large number of target genes, and we seek to answer two questions about this mode of regulation: what identifies the common target genes to the transcription factor; and how is the activity of the transcription factor itself regulated in response to intrinsic or extrinsic signals?

## Response elements identify genes under common regulation

The principle that emerges from characterizing groups of genes under common control is that *they share a promoter element that is recognized by a regulatory transcription factor*. An element that causes a gene to respond to such a factor is called a **response element**; examples are the HSE (heat shock response element), GRE (glucocorticoid response element), SRE (serum response element).

The properties of some inducible transcription factors and the elements that they recognize are summarized in Table 29.1. Response elements have the same general characteristics as upstream elements of promoters or enhancers. They contain short consensus sequences, and copies of the response elements found in different genes are closely related, but not necessarily identical. The region bound by the factor extends for a short distance on either side of

the consensus sequence. In promoters, the elements are not present at fixed distances from the startpoint, but are usually <200 bp upstream of it. The presence of a single element usually is sufficient to confer the regulatory response, but sometimes there are multiple copies.

Response elements may be located in promoters or in enhancers. Some types of elements are typically found in one rather than the other: usually an HSE is found in a promoter, while a GRE is found in an enhancer. We assume that all response elements function by the same general principle. *A gene is regulated by a sequence at the promoter or enhancer that is recognized by a specific protein. The protein functions as a transcription factor needed for RNA polymerase to initiate. Active protein is available only under conditions when the gene is to be expressed; its absence means that the promoter is not activated by this particular circuit.*

An example of a situation in which many genes are controlled by a single factor is provided by the heat shock response. This is common to a wide range of prokaryotes and eukaryotes and involves multiple controls of gene expression: an increase in temperature turns off transcription of some genes, turns on transcription of the **heat shock genes**, and causes changes in the translation of mRNAs. The control of the heat shock genes illustrates the differences between prokaryotic and eukaryotic modes of control. In bacteria, a new sigma factor is synthesized that directs RNA polymerase holoenzyme to recognize an alter-

**Table 29.1** Inducible transcription factors bind to response elements that identify groups of promoters or enhancers subject to coordinate control.

Regulatory Agent	Module	Consensus	Factor
Heat shock	HSE	CNNGAANNTCCNNG	HSTF
Glucocorticoid	GRE	TGGTACAAATGTTCT	Receptor
Phorbol ester	TRE	TGACTCA	AP1
Serum	SRE	CCATATTAGG	SRF

N. Leigh Anderson  
Ricardo Esquer-Blasco  
Jean-Paul Hofmann  
Norman G. Anderson

Large Scale Biology Corporation,  
Rockville, MD

## A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies

A standard two-dimensional (2-D) protein map of Fischer 344 rat liver (F344MST3) is presented, with a tabular listing of more than 1200 protein species. Sodium dodecyl sulfate (SDS) molecular mass and isoelectric point have been established, based on positions of numerous internal standards. This map has been used to connect and compare hundreds of 2-D gels of rat liver samples from a variety of studies, and forms the nucleus of an expanding database describing rat liver proteins and their regulation by various drugs and toxic agents. An example of such a study, involving regulation of cholesterol synthesis by cholesterol-lowering drugs and a high-cholesterol diet, is presented. Since the map has been obtained with a widely used and highly reproducible 2-D gel system (the Iso-Dalt<sup>®</sup> system), it can be directly related to an expanding body of work in other laboratories.

### Contents

1 Introduction	907
2 Material and methods	908
2.1 Sample preparation	908
2.2 Two-dimensional electrophoresis	909
2.3 Staining	909
2.4 Positional standardization	909
2.5 Computer analysis	909
2.6 Graphical data output	910
2.7 Experiment LSBC04	910
3 Results and discussion	910
3.1 The rat liver protein 2-D map	910
3.2 Carbamylated charge standards computed pI's and molecular mass standardization	911
3.3 An example of rat liver gene regulation: Cholesterol metabolism	911
3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated coordinately or inversely	911
3.3.2 MSN 235 and coregulated spots	912
3.3.3 An example of an anti-synergistic effect	912
3.3.4 Complexity of the cholesterol synthesis pathway	912
4 Conclusions	912
5 References	912
6 Addendum 1: Figures 1-13	914
7 Addendum 2: Tables 1-4	923
Table 1. Master table of proteins in rat liver database	923
Table 2. Table of some identified proteins	928
Table 3. Computed pI's of two sets of carbamylated protein standards: rabbit muscle CPK and human Hb	929
Table 4. Computed pI's of some known proteins related to measured CPK pI's	930

### 1 Introduction

High-resolution two-dimensional electrophoresis of proteins, introduced in 1975 by O'Farrell and others [1-4], has been used over the ensuing 16 years to examine a wide variety of biological systems, the results appearing in more than 5000 published papers. With the advent of computerized systems for analyzing two-dimensional (2-D) gel images and constructing spot databases, it is also possible to plan and assemble integrated bodies of information describing the appearance and regulation of thousands of protein gene products [5, 6]. Creating such databases involves amassing and organizing quantitative data from thousands of 2-D gels, and requires a substantial commitment in technology and resources.

Given the long-term effort required to develop a protein database, the choice of a biological system takes on considerable importance. While *in vitro* systems are ideal for answering many experimental questions, especially in cancer research and genetics, our experience with cell cultures and tissue samples suggests that some *in vivo* approaches could have major advantages. In particular, we have noticed that liver tissue samples from rats and mice appear to show greater quantitative reproducibility (in terms of individual protein expression) than replicate cell cultures. This is perhaps a natural result of the homeostasis maintained in a complete animal vs. the well-known variability of cell cultures, the latter due principally to differences in reagents (e.g., fetal bovine serum), conditions (e.g., pH) and genetic "evolution" of cell lines while in culture. It is also more difficult to generate adequate amounts of protein from cell culture systems (particularly with attached cells), forcing the investigator to resort to radioisotope-based or silver-based stain-detection methods. While these methods are more sensitive (sometimes much more sensitive) than the Coomassie Brilliant Blue (CBB) stain typically used for protein detection in "large" protein samples, they are generally more variable, more labor-intensive and, in the case of radiographic methods, may generate highly "noisy" images, due to the properties of the films used. By contrast, large protein samples can easily be prepared from liver using urea/Nonidet P-40 (NP-40) solubilization and stained with CBB, which has the advantage of being easily reproducible [8]. Finally, there remains the question of the "truthfulness" of many *in vitro* systems as compared to their *in vivo* analogs; how great are the changes caused by the introduction into a cul-

Correspondence: Dr. N. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850, USA

Abbreviations: CBB, Coomassie Brilliant Blue; CPK, creatine phosphokinase; 2-D, two-dimensional; IEF, isoelectric focusing; MSN, master spot number; NP-40, Nonidet P-40; SDS, sodium dodecyl sulfate

ture and the associated shift to strong selection for growth, and how do these affect experimental outcomes? Hence the apparent advantages of *in vitro* systems, in terms of experimental manipulation, may be counterbalanced by other factors relating to 2-D data quality.

There is a second important class of reasons for exploring the use of an *in vivo* biological system such as the liver. Historically, there have been two broad approaches to the mechanistic dissection of biochemical processes in intact cellular systems: genetics (a search for informative mutants) and the use of chemical agents (drugs and chemical toxins). Both approaches help us to understand complex systems by disrupting some specific functional element and showing us the result. With the development of techniques for genetic manipulation and cloning, the genetic approach can be effectively applied either *in vitro* or *in vivo*, although the *in vitro* route is usually quicker. The chemical approach can also be applied to either sort of biological system; here, however, the bulk of consistently acquired information is in experimental animals (rats and mice). While most biologists know a short list of compounds having specific, experimentally useful effects (e.g., inhibitors of protein synthesis, ionophores, polymerase inhibitors, channel blockers, nucleotide analogs, and compounds affecting polymerization of cytoskeletal proteins), there is a much larger number of interesting chemically-induced effects, most of them characterized by toxicologists and pharmacologists in rodent systems. Just as a thorough genetic analysis would involve saturating a genome with mutations, it is possible to imagine a saturating number of drugs, the analysis of whose actions would reveal the complete biochemistry of the cell. While organized drug discovery efforts usually target specific desired effects, the nature of the process, with its dependence on screening large numbers of compounds, necessarily produces many unanticipated effects. It is therefore reasonable to suppose that the required broad range of compounds necessary to achieve "biochemical saturation" may be forthcoming; in fact, it may already exist among the hundreds of thousands of compounds that failed to qualify as drugs.

Among organs, the liver is an obvious choice for the study of chemical effects because of its well-known plasticity and responsiveness. The brain appears to be quite plastic (e.g. [7]), but it is a complicated mixture of cell types requiring skillful dissection for most experiments. The kidney, while quite responsive, also presents a potentially confounding mixture of cell types. The liver, by contrast, is made up of one predominant cell type which is easy to solubilize: the hepatocyte, representing more than 95% of its mass. Most importantly, the liver performs many homeostatic functions that require rapid modulation of gene expression. It appears that most chemical agents tested affect gene expression in the liver at some dosage (N. Leigh Anderson, unpublished observations), an interesting contrast to our earlier work with lymphocytes, for example, which seem to be much less responsive. Such results conform to the expectation that cells with a homeostatic, physiological role should be more plastic than cells differentiated for a purpose dependent on the action of a limited number of specific genes.

The liver also allows the parallels between *in vitro* and *in vivo* systems to be examined in detail. Significant progress

has been made in the development of mouse, rat and human hepatocyte culture systems, as well as in precision-cut tissue slices. Using such an array of techniques, it is possible to assemble a matrix of mammalian systems including mouse and rat *in vivo* on one level and mouse, rat and human *in vitro* on a second level, and to compare effects between species and between systems. This approach allows us to draw informed conclusions regarding the biochemical "universality" of biological responses among the mammals and to offer some insight into the validity of *in vitro* approaches for toxicological screening. We believe this data will be necessary if *in vitro* alternatives are to achieve wide usage in government-mandated safety testing of drugs, consumer products and industrial and agricultural chemicals.

A number of interesting studies have been published using 2-D mapping to examine effects in the rodent liver. A number of investigators have made use of the technique to screen for existing genetic variants [8-11] or induced mutations [12-14], mainly in the mouse. This work builds on the wealth of genetic information available on the mouse and its established position as a mammalian mutation-detection system. While some studies of chemical effects have been undertaken in the mouse [15-17], most have used the rat [18-23]. The examination of the cytochrome p-450 system, in particular, has been carried out almost exclusively on the rat [24, 25].

These considerations lead us to conclude that rodent liver offers the best opportunity to systematically examine an array of gene regulation systems, and ultimately to build a predictive model of large-scale mammalian gene control. The basic underlying foundation of such a project is a reliable, reproducible master 2-D pattern of liver, to which ongoing experimental results can be referred. In this paper, we report such a master pattern for the acidic and neutral proteins of rat liver (pattern F344MST3). In future, this master will be supplemented by maps of basic proteins, and analogous maps of mouse and human liver.

## 2 Materials and methods

### 2.1 Sample preparation

Liver is an ideal sample material for most biochemical studies, including 2-D analysis. A sample is taken of approximately 0.5 g of tissue from the apical end of the left lobe of the liver. Solubilization is effected as rapidly as practical; a delay of 5-15 min appears to cause no major alteration in liver protein composition if the liver pieces are kept cold (e.g., on ice) in the interim. In the solubilization process, the liver sample is weighed, placed in a glass homogenizer (e.g., 15 mL Wheaton); 8 volumes of solubilizing solution\*

\* The solubilizing solution is composed of 2% NP-40 (Sigma), 9 M urea (analytical grade, e.g., BDH or Bio-Rad), 0.5% dithiothreitol (DTT; Sigma) and 2% carrier ampholytes (pH 9-11 LKB: these come as a 20% stock solution, so 2% final concentration is achieved by making the final solution 10% 9-11 Ampholine by volume). A large batch of solubilizer (several hundred mL) is made and stored frozen at -80°C in aliquots sufficient to provide enough for one day's estimated sample preparation requirement. The solution is never allowed to become warmer than room temperature at any stage during preparation or thawing for use, since heating of concentrated urea solutions can produce contaminants that covalently modify proteins producing artifactual charge shifts. Once thawed, any unused solubilizer is discarded.

added (i.e., 4 mL per 0.5 g tissue) and the mixture is homogenized using first the loose- and then the tight-fitting glass pestle. This takes approximately 5 strokes with the pestle and is carried out at room temperature because urea would crystallize out in the cold. Once the liver sample is thoroughly homogenized in the solubilizer, it is assumed that all the proteins are denatured (by the chaotropic effect of the urea and NP-40 detergent) and the enzymes inactivated by the high pH (~9.5). Therefore these samples may be kept at room temperature until they can be centrifuged frozen as a group (within several hours of preparation). The samples are centrifuged for  $6 \times 10^5$  g min (e.g., 500 000 g for 12 min using a Beckman TL-100 centrifuge). The centrifuge rotor is maintained at just below room temperature (e.g., 15–20°C), but not too cold, so as to prevent the precipitation of urea. The centrifuge of choice is a Beckman TL-100 because of the sample tube sizes available, but any ultracentrifuge accepting smallish tubes will suffice. When an appropriate centrifuge is not available near the site of sample preparation, samples can be frozen at –80°C and thawed prior to centrifugation and collection of supernatants. Each supernatant is carefully removed following centrifugation and aliquoted into at least 4 clean tubes for storage. This is done by transferring all the supernatant to one clean tube, mixing this gently (to assure homogeneous composition) and then dividing it into 4 aliquots. The aliquots are frozen immediately at –80°C. These multiple aliquots can provide insurance against a failed run or a freezer breakdown.

## 2. Two-dimensional electrophoresis

Sample proteins are resolved by 2-D electrophoresis using the 20 × 25 cm Iso-Dalt\* 2-D gel system ([26–29]; produced by LSB and by Hoefer Scientific Instruments, San Francisco) operating with 20 gels per batch. All first-dimensional isoelectric focusing (IEF) gels are prepared using the same single standardized batch of carrier ampholytes (BDH 4–8A in the present case, selected by LSB's batch-testing program for rat and mouse database work\*\*). A 10 µL sample of solubilized liver protein is applied to each gel, and the gels are run for 33 000 to 34 500 volt-hours using a progressively increasing voltage protocol implemented by a programmable high-voltage power supply. An Angeliq™ computer-controlled gradient-casting system (produced by LSB) is used to prepare second-dimensional sodium dodecyl sulfate (SDS) polyacrylamide gradient slab gels in which the top 5% of the gel is 11%T acrylamide, and the lower 95% of the gel varies linearly from 11% to 18%T.

This system has recently been modified so as to employ a commercially available 30.8%T acrylamide/*N,N*-methylenebisacrylamide prepared solution (thus avoiding the handling of the solid acrylamide monomer) and three additional stock solutions: buffer (made from Sigma pre-set Tris), persulfate and *N,N,N,N*-tetramethylethylenediamine (TEMED). Each gel is identified by a computer-printed filter paper label polymerized into the lower left corner of the gel. First-dimensional IEF tube gels are loaded

directly (as extruded) onto the slab gels without equilibration, and held in place by polyester fabric wedges (Wedgies™, produced by LSB) to avoid the use of hot agarose. Second-dimensional slab gels are run overnight, in groups of 20, in cooled DALT tanks (10°C) with buffer circulation. All run parameters, reagent source and lot information, and notations of deviation from expected results are entered by the technician responsible on a detailed, multi-page record of the experiment.

## 2.3 Staining

Following SDS-electrophoresis, slab gels are stained for protein using a colloidal Coomassie Blue G-250 procedure in covered plastic boxes, with 10 gels (totalling approximately 1 L of gel) per box. This procedure (based on the work of Neuhoff [30, 31]) involves fixation in 1.5 L of 50% ethanol and 2% phosphoric acid for 2 h, three 30 min washes, each in 2 L of cold tap water, and transfer to 1.5 L of 34% methanol, 17% ammonium sulfate and 2% phosphoric acid for 1 h, followed by the addition of a gram of powdered Coomassie Blue G-250 stain. Staining requires approximately 4 days to reach equilibrium intensity, whereupon gels are transferred to cool tap water and their surfaces rinsed to remove any particulate stain prior to scanning. Gels may be kept for several months in water with added sodium azide. The water washes remove ethanol that would dissolve the stain (and render the system noncolloidal, with high backgrounds). The concentrated ammonium sulfate and methanol solution is diluted by equilibration with the water volume of the gels to automatically achieve the correct final concentrations for colloidal staining. Practical advantages of this staining approach can be summarized as follows: (i) the low, flat background makes computer evaluation of small spots (max OD < 0.02) possible, especially when using laser densitometry; (ii) up to 1500 spots can be reliably detected on many gels (e.g., rat liver) at loadings low enough to preserve excellent resolution; and (iii) reproducibility appears to be very good: at least several hundred spots have coefficients of reproducibility less than 15%. This value is at least as good as previous CBB methods, and significantly better than many silver stain systems.

## 2.4 Positional standardization

The carbamylated rabbit muscle creatine phosphokinase (CPK) standards [32] are purchased from Pharmacia and BDH. Amino acid compositions, and numbers of residues present in proteins used for internal standardization, are taken from the Protein Identification Resource (PIR) sequence database [33].

## 2.5 Computer analysis

Stained slab gels are digitized in red light at 134 micron resolution, using either a Molecular Dynamics laser scanner (with pixel sampling) or an Eikonix 78/99 CCD scanner. Raw digitized gel images are archived on high-density DAT tape (or equivalent storage media) and a greyscale video-print prepared from the raw digital image as hard-copy backup of the gel image. Gels are processed using the Kepler® software system (produced by LSB), a commercially available workstation-based software package built on

\*\*This material (succeeding certified batches of which are available from Hoefer Scientific Instruments) has the most linear pH gradient produced by any ampholyte tested except for the Pharmacia wide range (which has an unacceptable tendency to bind high-molecular weight acidic proteins, causing them to streak).

some of the principles of the earlier TYCHO system [34-41]. Procedure PROC008 is used to yield a spotlist giving position, shape and density information for each detected spot. This procedure makes use of digital filtering, mathematical morphology techniques and digital masking to remove the background, and uses full 2-D least-squares optimization to refine the parameters of a 2-D Gaussian shape for each spot. Processing parameters and file locations are stored in a relational database, while various log files detailing operation of the automatic analysis software are archived with the reduced data. The computed resolution and level of Gaussian convergence of each gel are inspected and archived for quality control purposes.

Experiment packages are constructed using the Kepler experiment definition database to assemble groups of 2-D patterns corresponding to the experimental groups (e.g., treated and control animals). Each 2-D pattern is matched to the appropriate "master" 2-D pattern (pattern F344MST3 in the case of Fischer 344 rat liver), thereby providing linkage to the existing rodent protein 2-D databases. The software allows experiments containing hundreds of gels to be constructed and analyzed as a unit, with up to 100 gels displayed on the screen at one time for comparative purposes and multiple pages to accommodate experiments of > 1000 gels. For each treatment, proteins showing significant quantitative differences vs. appropriate controls are selected using group-wise statistical parameters (e.g., Student's *t*-test, Kepler<sup>®</sup> procedure STUDENT). Proteins satisfying various quantitative criteria (such as  $P < 0.001$  difference from appropriate controls) are represented as highlighted spots onscreen or on computer-plotted protein maps and stored as spot populations (i.e., logical vectors) in a liver protein database. Quantitative data (spot parameters, statistical or other computed values) are stored as real-valued vectors in the database. Analysis of coregulation is performed using a Pierson product-moment correlation (Kepler procedure CORREL) to determine whether groups of proteins are coordinately regulated by any of the treatments. Such groups can be presented graphically on a protein map, and reported together with the statistical criteria used to assess the level of coregulation. Multivariate statistical analysis (e.g., principal components' analysis) is performed on data exported to SAS (SAS Institute).

## 2.6 Graphical data output

Graphical results are prepared in GKS and translated within Kepler<sup>®</sup> into output for any of a variety of devices. Linedrawing output is typically prepared as Postscript and printed on an Apple Laserwriter. Detailed maps presented here have been generated using an ultra-high-resolution Postscript-compatible Linotronic output device. Greyscale graphics are reproduced from the workstation screen using a Seikosha videoprinter. Patterns are shown in the standard orientation, with high molecular mass at the top and acidic proteins to the left.

## 2.7 Experiment LSBC04

In the study described here 12-week-old Charles River male F344 rats were used. Diets were prepared at LSB, based on a Purina 5755M Basal Purified Diet. Lovastatin and cholestyramine were obtained as prescription pharma-

ceuticals, ground and mixed with the diet at concentrations of 0.075% and 1%, respectively. The high cholesterol diet was Purina 5801M-A (5% cholesterol plus 1% sodium cholate in the control diet). Animal work was carried out by Microbiological Associates (Bethesda, MD). Animals were acclimatized for one week on the control diet, fed test or control diets for one week, and sacrificed on day 8. Average daily doses of lovastatin and cholestyramine in appropriate groups were 37 mg/kg/day and 5 g/kg/day, respectively, based on the weight of the food consumed. Liver samples were collected and prepared for 2-D electrophoresis according to the standard liver protocol (homogenization in 8 volumes of 9 M urea, 2% NP-40, 0.5% dithiothreitol, 2% LKB pH 9-11 carrier ampholytes, followed by centrifugation for 30 min at 80 000  $\times$  g). Kidney, brain and plasma samples were frozen. Gels were run as described above, and the data was analyzed using the Kepler<sup>®</sup> system. Gels were scaled, to remove the effect of differences in protein loading, by setting the summed abundances of a large number of matched spots equal for each gel (linear scaling).

## 3 Results and discussion

### 3.1 The rat liver protein 2-D map

F344MST3 is a standard 2-D pattern of rat liver proteins, based on the Fischer 344 strain. This pattern was initiated from a single 2-D gel and extensively edited in an experiment comparing it to a range of protein loads, so as to include both small spots and well-resolved representations of high-abundance spots. More than 700 rat liver 2-D patterns have been matched to F344MST3 in a series of drug effects and protein characterization experiments, and numerous new spots (induced by specific drugs, for instance) have been added as a result. A modified version including additional spots present in the Sprague-Dawley outbred rat has also been developed (data not shown). Figure 1 shows a greyscale representation and Fig. 2 a schematic plot of the master pattern. More than 1200 spots are included, most of which are visible on typical gels loaded with 10  $\mu$ L of solubilized liver protein prepared by the standard method and stained with colloidal Coomassie Blue. Master spot numbers (MSN's) have been assigned to all proteins, and appear in the following figures, each showing one quadrant of the pattern. Figure 3 shows the upper left (acidic, high molecular mass) quadrant, Fig. 4 the upper right (basic, high molecular mass) quadrant, Fig. 5 the lower left (acidic, low molecular mass) quadrant, and Fig. 6 the lower right (basic, low molecular mass) quadrant. The quadrants overlap as an aid to moving between them. The gel position (in 100 micron units), isoelectric point (relative to the CPK internal pI standards) and SDS molecular mass (from the calibration curve in Fig. 8) are listed for each spot (Table 1). Because of the precision of the CPK-pI values, these parameters can be used to relate spot locations between gel systems more reliably than using pI measurements expressed as pH. A major objective of current studies is the identification of all major spots corresponding to known liver proteins, as well as rigorous definitions of subcellular organelle contents. Of particular interest to us is the parallel development of identifications in the rat and mouse liver maps, allowing detailed comparisons of gene expression effects in the two systems. The results of these studies will be presented systematically in a later edition of this database.

We include here a useful series of 22 orienting identifications as an aid to other users of the rat liver pattern (Table 1).

## 2. Carbamylated charge standards, computed pI's and molecular mass standardization

We have previously shown that the use of a system of close-spaced internal pI markers (made by carbamylating a basic protein) offers an accurate and workable solution to the problem of assigning positions in the pI dimension [32]. The same system, based on 36 protein species made by carbamylating rabbit muscle CPK, has been used here to assign pI's to most rat liver acidic and neutral proteins. The standards were coelectrophoresed with total liver proteins, and the standard spots added to a special version of the master pattern F344MST3. The gel X-coordinates of all liver protein spots lying within the CPK charge train were then transformed into CPK pI positions by interpolation between the positions of immediately adjacent standards (Table 1) using a Kepler<sup>2</sup> vector procedure.

It has proven possible to compute fairly accurate pI values for many proteins from the amino acid composition [42]. We have attempted here to test a further elaboration of this approach, in which we computed pI's for the CPK standards themselves, based on our knowledge of the rabbit muscle CPK sequence and the fact that adjacent members of the charge train typically differ by blockage of one additional lysine residue (Table 3). We compared these values to similar computed pI's for an additional set of carbamylated standards made from human hemoglobin beta chains and a series of rat liver and human plasma proteins of known position and sequence (Fig. 7, Table 4). The result demonstrates good concordance between these systems. Two proteins show significant deviations: liver fatty-acid binding protein (FABP; #1 in Table 4) and protein disulphide isomerase (#20 in the table). The FABP spot present on F344MST3 may represent a charge-modified version of a more basic parent spot closer to the expected pI, not resolved in the IEF/SDS gel. Of particular importance is the fact that, by comparing computed pI's of sequenced but unlocated proteins with the CPK pI's, we can assign a probable gel location without making any assumptions regarding the actual gel pH gradient. This offers a useful shortcut, given the vagaries of pH measurement on small diameter IEF gels. We have used this approach to compute the CPK pI's of all rat and mouse proteins in the PIR sequence database, as an aid to protein identification (data not shown).

In order to standardize SDS molecular weight (SDS-MW), we have used a standard curve fitted to a series of identified proteins (Fig. 8). Rather than using molecular mass *per se*, we have elected to use the number of amino acids in the polypeptide chain, as perhaps a better indication of the length of the SDS-coated rod that is sieved by the second dimension slab. The resulting values were multiplied by 12 (the weighted average mass of amino acids in sequenced proteins) to give predicted molecular masses. Because we use gradient slabs, we have not constrained the fitted curve to conform to any predetermined model; rather we tried many equations and selected the best using the program "Tablecurve" on a PC. The equation chosen was  $y = a + bx + c/x^2$ , where  $y$  is the number of residues,  $x$  is the gel

Y coordinate,  $a$  is 511.83,  $b$  is -0.2731 and  $c$  is 33183801. The resulting fit appears to be fairly good over a broad range of molecular mass.

## 3.3 An example of rat liver gene regulation: Cholesterol metabolism

Experiment LSBC04 was designed as a small-scale test of the regulation of cholesterol metabolism *in vivo* by three agents included in the diet: lovastatin (Mevacor<sup>®</sup>, an inhibitor of HMG-CoA reductase); cholestyramine (a bile acid sequestrant that has the effect of removing cholesterol from the gut-liver recirculation); and cholesterol itself. The first two agents should lower available cholesterol and the third should raise it, allowing manipulation of relevant gene expression control systems in both directions. Such an experiment offers an interesting test of the 2-D mapping system since most of the pathway enzymes are present in low abundance, many are membrane-bound and difficult to solubilize, and the pathway itself is complex. Approximately 1000 proteins were separated and detected in liver homogenates. Twenty-one proteins were found to be affected by at least one treatment, and these could be divided into several coregulated groups.

### 3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated coordinately or inversely

One group of spots (including a spot assigned to the cytosolic HMG-CoA synthase, MSN 413) showed the expected increase in abundance with lovastatin or cholestyramine, the synergistic further increase with lovastatin and cholestyramine, and a dramatic decrease with the high cholesterol diet. Spot number 413 is the most strongly regulated protein in the present experiment, showing a 5- to 10-fold induction after a 1 week treatment with 0.075% lovastatin and 1% cholestyramine in the diet (Figs. 9 and 10). Its expression follows precisely the expectation for an enzyme whose abundance is controlled by the cholesterol level; it is progressively increased from the control levels by cholestyramine, lovastatin and lovastatin plus cholestyramine, and it sinks below the threshold of detection in animals fed the high cholesterol diet. This spot has been tentatively identified as the cytosolic HMG-CoA synthase, based on a reaction with an antiserum to that protein provided by Dr. Michael Greenspan at Merck Sharp & Dohme Research Laboratories. This enzyme lies immediately before HMG-CoA reductase in the liver cholesterol biosynthesis pathway, and is known to be co-regulated with it. Spot 413 has an SDS molecular weight of about 54 000 and a CPK pI of -11.4, in reasonably close agreement with a molecular weight of 57 300 and a CPK pI of -15.7 computed from the known sequence of the hamster enzyme [43].

Using a classical product-moment correlation test (Kepler procedure CORREL), a series of five additional spots was found to be coregulated with 413. The level of correlation was exceedingly high (> 95%). Two of these, 1250 and 933, are at similar molecular weights and approximately one charge more acidic than 413 (Fig. 9), indicating that they may be covalently modified forms of the 413 polypeptide. This suspicion is strengthened by the observation that both spots are also stained by the antibody to cytosolic HMG-CoA synthase. The remaining three correlated spots appear

to comprise an additional related pair (1253 and 1001) of around 40 kDa and a single spot (1119) of around 28 kDa. Because these two presumed proteins are present at substantially lower abundances than 413, and because the cytosolic HMG-CoA synthase is reported to consist of only one type of polypeptide, they are likely to represent other, very tightly coregulated enzymes. A second group of six spots was selected based on a regulatory pattern close to the inverse of that for spot 413 (MSN's 34, 79, 178, 182, 204, 347; data not shown). For these proteins, the lowest level of expression occurs with exposure to lovastatin plus cholestyramine and the highest level upon exposure to the high-cholesterol diet. Spots 182 and 79 are highly correlated and lie about one charge apart at the same molecular weight; they may thus be isoforms of a single protein. The other four spots probably represent additional enzymes or subunits.

### 3.3.2 MSN 235 and coregulated spots

A third group of five spots, mainly comprised of mitochondrial proteins including putative mitochondrial HMG-CoA synthase spots, showed a modest induction by lovastatin alone, but little or no effect with any of the other treatments (including the combination of lovastatin and cholestyramine; Fig. 12). This result is intriguing because lovastatin was expected to affect only the regulation of enzymes of cholesterol synthesis, which is entirely extra-mitochondrial. Three of the spots (235, 134, 144) form a closely-packed triad at approximately 30 kDa, and are likely to represent isoforms of one protein. All three spots are stained by an antibody to the mitochondrial form of HMG-CoA synthase obtained from Dr. Greenspan. Subcellular fractionation indicates a mitochondrial location. The other two spots (633 at about 38 kDa and 724 at about 69 kDa) are each present at lower abundance than the members of the triad.

### 3.3.3 An example of an anti-synergistic effect

A sixth spot (367) shows strong induction by lovastatin (two- to threefold), and about half as much induction with lovastatin plus cholestyramine, but without sharing the animal-animal heterogeneity pattern of the 235-set (Fig. 13). This protein is also mitochondrial, and represents the clearest example of an anti-synergistic effect of lovastatin and cholestyramine. The existence of such an effect demonstrates that lovastatin and cholestyramine do not act exclusively through the same regulatory pathway.

### 3.3.4 Complexity of the cholesterol synthesis pathway

Taken together, these results suggest that treatment with lovastatin alone can affect both cytosolic and mitochondrial pathways using HMG-CoA, while cholestyramine, on the other hand, either alone or in combination with lovastatin, produces a strong effect on the putative cytosolic pathway, but little or no effect on the putative mitochondrial pathway. An explanation for this difference may lie in lovastatin's effect on levels of HMG-CoA and related precursor compounds that are exchanged between the cytosol and the mitochondrion, whereas cholestyramine should affect only the cytosolic pathways directly controlled by cholesterol and bile acid levels. It remains to be explained why some

proteins of the putative mitochondrial pathway are so much more variable in their expression in all groups. An examination of all the coregulated groups suggests that quantitative statistical techniques can extract a wealth of interesting information from large sets of reproducible gels. The abundance of spots in the 413 coregulation group, for example, shows an amazing level of concordance in their relative expression among the five individuals of the lovastatin and cholestyramine treatment group. This effect is not due to differences in total protein loading, since they have already been removed by scaling, and since proteins with quite different regulation patterns can be demonstrated (e.g., Fig. 13). Such effects raise the possibility that many gene coregulation sets may be revealed through the study of a sufficiently large population of control animals (*i.e.*, without any experimental manipulation). This approach, exploiting natural biological variation in protein expression instead of drug effects, offers an important incentive for the construction of a large library of control animal patterns.

## 4 Conclusions

Because of the widespread use of rat liver in both basic biochemistry and in toxicology, there is a long-term need for a comprehensive database of liver proteins. The rat liver master pattern presented here has proven to be an accurate representation of this system, having been matched to more than 700 gels to date. As the number of proteins identified and the number of compounds tested for gene expression effects grows, we expect this database to contribute valuable insights into gene regulation. Its practical utility in several areas of mechanistic toxicology is already being demonstrated.

Received September 11, 1991

## 5 References

- [1] O'Farrell, P., *J. Biol. Chem.* 1975, 250, 4007-4021.
- [2] Klose, J., *Humangenetik* 1975, 26, 231-243.
- [3] Scheele, G. A., *J. Biol. Chem.* 1975, 250, 5375-5385.
- [4] Iborra, G. and Buhler, J. M., *Anal. Biochem.* 1976, 74, 503-511.
- [5] Anderson, N. G. and Anderson, N. L., *Behring. Inst. Mitt.* 1979, 63, 169-210.
- [6] Anderson, N. G. and Anderson, N. L., *Clin. Chem.* 1982, 28, 739-746.
- [7] Heydorn, W. E., Creed, G. J. and Jacobowitz, D. M., *J. Pharmacol. Exp. Therap.* 1984, 229, 622-628.
- [8] Anderson, N. L., Nance, S. L., Tollaksen, S. L., Giere, F. A. and Anderson, N. G., *Electrophoresis* 1985, 6, 592-599.
- [9] Racine, R. R. and Langley, C. H., *Biochem. Genet.* 1980, 18, 185-197.
- [10] Klose, J., *Mol. Evol.* 1982, 18, 315-328.
- [11] Neel, J. V., Baier, L., Hanash, S. and Erickson, R. P., *J. Hered.* 1985, 76, 314-320.
- [12] Marshall, R. R., Raj, A. S., Grant, F. J. and Heddle, J. A., *Can. J. Genet. Cytol.* 1983, 25, 457-446.
- [13] Taylor, J., Anderson, N. L., Anderson, N. G., Gemmell, A., Giometti, C. S., Nance, S. L. and Tollaksen, S. L., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, Verlag Chemie, Weinheim 1986, pp. 583-587.
- [14] Giometti, C. S., Gemmell, M. A., Nance, S. L., Tollaksen, S. L. and Taylor, J., *J. Biol. Chem.* 1987, 262, 12764-12767.
- [15] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., in: Galteau, M.-M. and Siest, G. (Eds.), *Progrès Récents en Electrophorèse Bidimensionnelle*, Presses Universitaires de Nancy, Nancy 1986, pp. 253-260.
- [16] Anderson, N. L., Swanson, M., Giere, F. A., Tollaksen, S., Gemmell, A., Nance, S. L. and Anderson, N. G., *Electrophoresis* 1986, 7, 44-48.



- Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., *Fundam. Appl. Toxicol.* 1987, 8, 39-50.
- Anderson, N. L., in: *New Horizons in Toxicology*, Eli Lilly Symposium, 1991, in press.
- Antoine, B., Rahimi-Pour, A., Sies, G., Magdalou, J. and Galteau, M. M., *Cell. Biochem. Funct.* 1987, 5, 217-231.
- Elliott, B. M., Ramasamy, R., Stonard, M. D. and Spragg, S. P., *Biochim. Biophys. Acta* 1986, 876, 135-140.
- Huber, B. E., Heilman, C. A., Wirth, P. J., Miller, M. J. and Thorgeirsson, S. S., *Hepatology* 1986, 6, 206-219.
- Wirth, P. J. and Vesterberg, O., *Electrophoresis* 1988, 9, 47-53.
- Witzmann, F. A. and Parker, D. N., *Toxicol. Lett.* 1991, 57, 29-36.
- Rampersaud, A., Waxman, D. J., Ryan, D. E., Levin, W. and Walz, F. G., Jr., *Arch. Biochem. Biophys.* 1985, 243, 174-183.
- Vlasuk, G. P. and Walz, F. G., Jr., *Anal. Biochem.* 1980, 105, 112-120.
- Anderson, N. G. and Anderson, N. L., *Anal. Biochem.* 1978, 85, 331-340.
- Anderson, N. L. and Anderson, N. G., *Anal. Biochem.* 1978, 85, 341-354.
- Anderson, L., Hofmann, J.-P., Anderson, E., Walker, B. and Anderson, N. G., in: Endler, A. T. and Hanash, S. (Eds.), *Two-Dimensional Electrophoresis*, VCH Verlagsgesellschaft, Weinheim 1989, pp. 288-297.
- Anderson, L., *Two-Dimensional Electrophoresis: Operation of the ISO-DALT<sup>®</sup> System*, Large Scale Biology Press, Washington, DC 1988, ISBN 0-945532-00-8, 170pp.
- Neuhoff, V., Stamm, R. and Eibl, H., *Electrophoresis* 1985, 6, 427-448.
- [31] Neuhoff, V., Arold, N., Taube, D. and Ehrhardt, W., *Electrophoresis* 1988, 9, 255-262.
- [32] Anderson, N. L. and Hickman, B. J., *Anal. Biochem.* 1979, 93, 312-320.
- [33] Sidman, K. E., George, D. E., Barker, W. C. and Hunt, L. T., *Nucl. Acids Res.* 1988, 16, 1869-1871.
- [34] Taylor, J., Anderson, N. L., Coulter, B. P., Scandora, A. E. and Anderson, N. G., in: Radola, B. J. (Ed.), *Electrophoresis '79*, de Gruyter, Berlin 1980, pp. 329-339.
- [35] Taylor, J., Anderson, N. L. and Anderson, N. G., in: Allen, R. C. and Arnaud, P. (Eds.), *Electrophoresis '81*, de Gruyter, Berlin 1981, pp. 383-400.
- [36] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. and Anderson, N. G., *Clin. Chem.* 1981, 27, 1807-1820.
- [37] Taylor, J., Anderson, N. L., Scandora, A. E., Jr., Willard, K. E. and Anderson, N. G., *Clin. Chem.* 1982, 28, 861-866.
- [38] Taylor, J., Anderson, N. L. and Anderson, N. G., *Electrophoresis* 1983, 4, 338-345.
- [39] Anderson, N. L. and Taylor, J., in: *Proceedings of the Fourth Annual Conference and Exposition of the National Computer Graphics Association*, Chicago, June 26-30, 1983, pp. 69-76.
- [40] Anderson, N. L., Hofmann, J.-P., Gemmell, A. and Taylor, J., *Clin. Chem.* 1984, 30, 2031-2036.
- [41] Anderson, L., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '88*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 313-321.
- [42] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E. and Phillips, T. A., *Electrophoresis* 1989, 10, 116-121.
- [43] Gil, G., Goldstein, J. L., Slaughter, C. A. and Brown, M. S., *J. Biol. Chem.* 1986, 261, 3710-3716.



## 6 Addendum 1: Figures 1-13

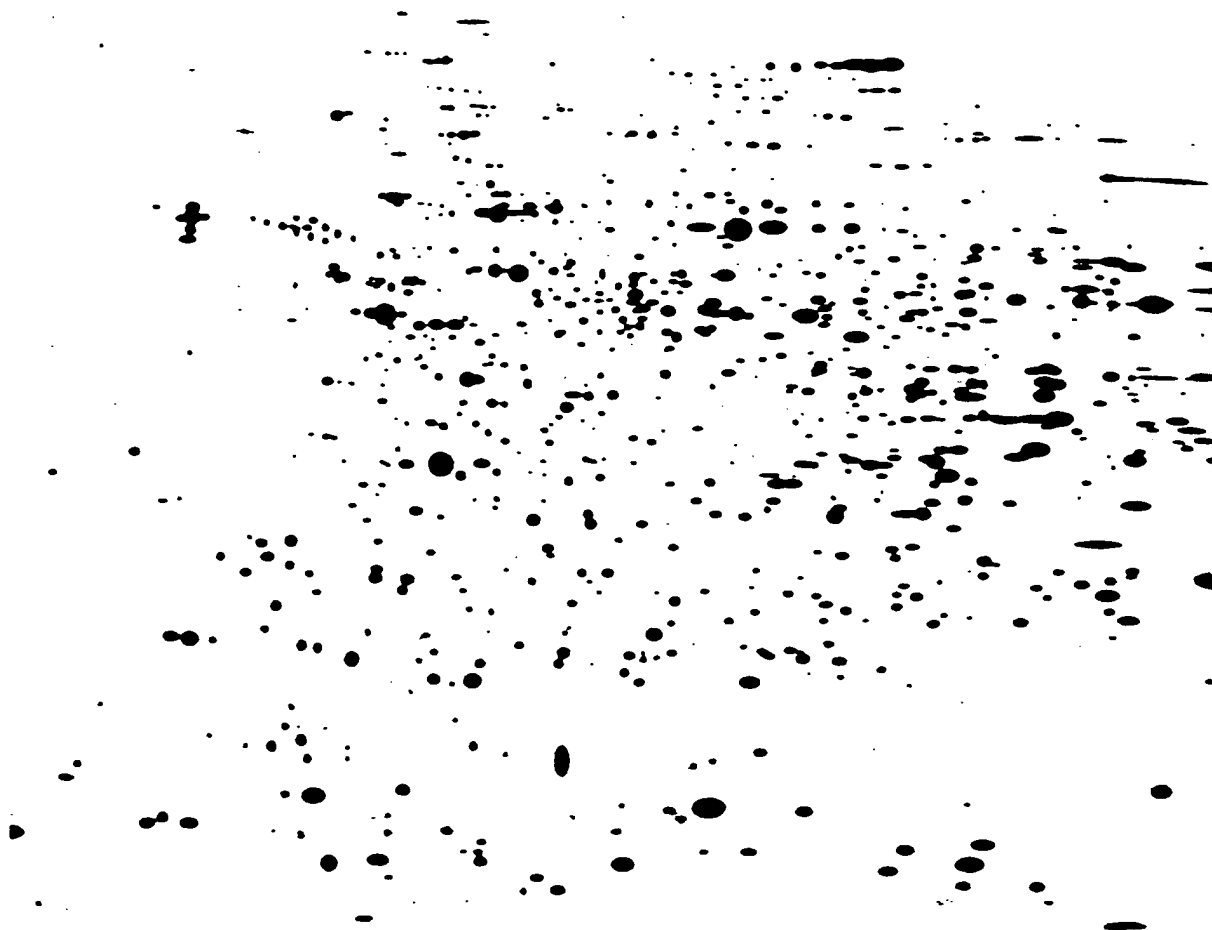


Figure 1. Synthetic representation of the standard rat liver 2-D master pattern, rendered as a greyscale image using a videoprinter.

Fig. 2. Schem  
ants.

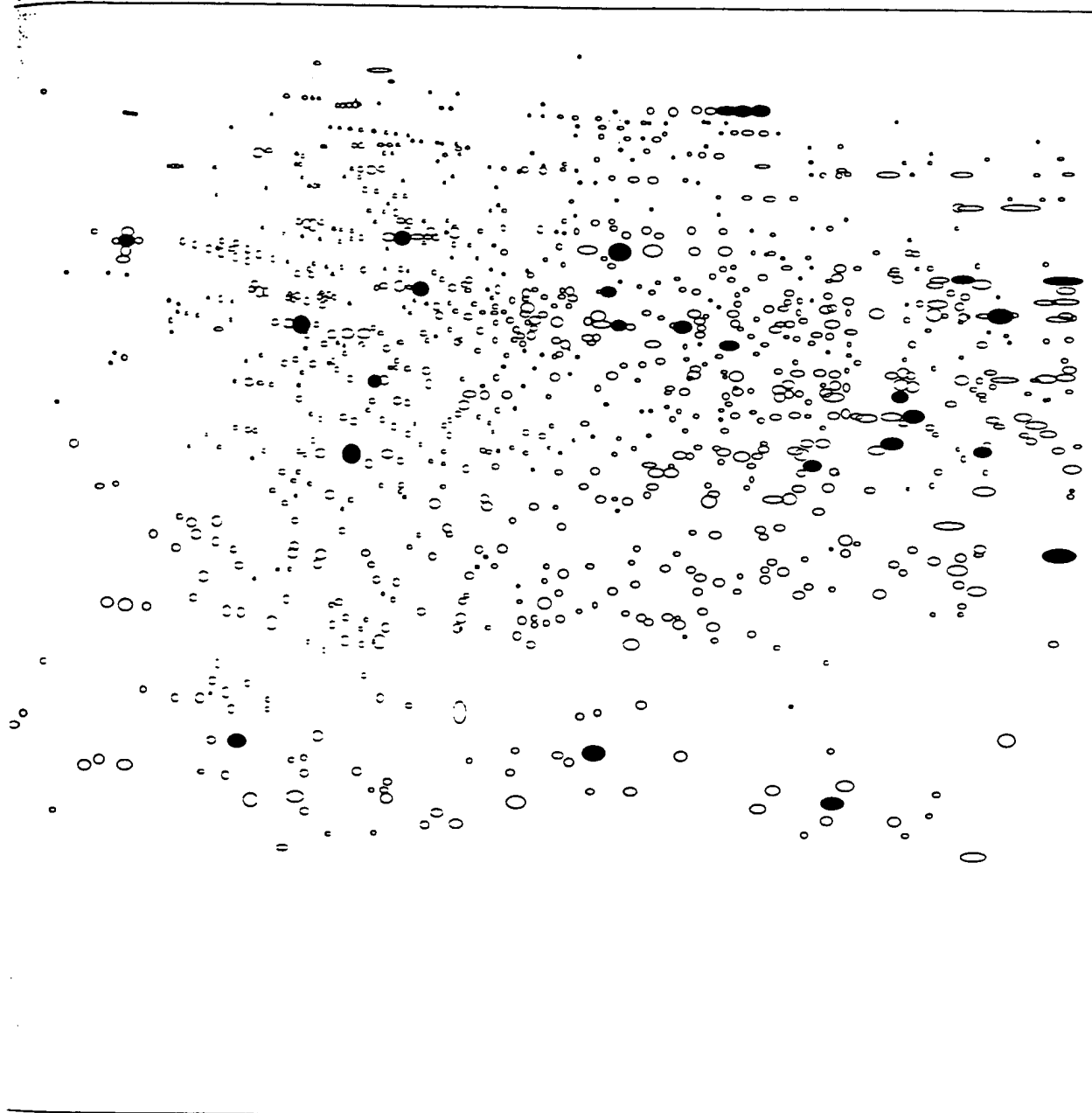


Fig. 2. Schematic representation of the master pattern (the same as Fig. 1), useful as an aid in relating specific areas of Fig. 1 and the following detailed prints.

1

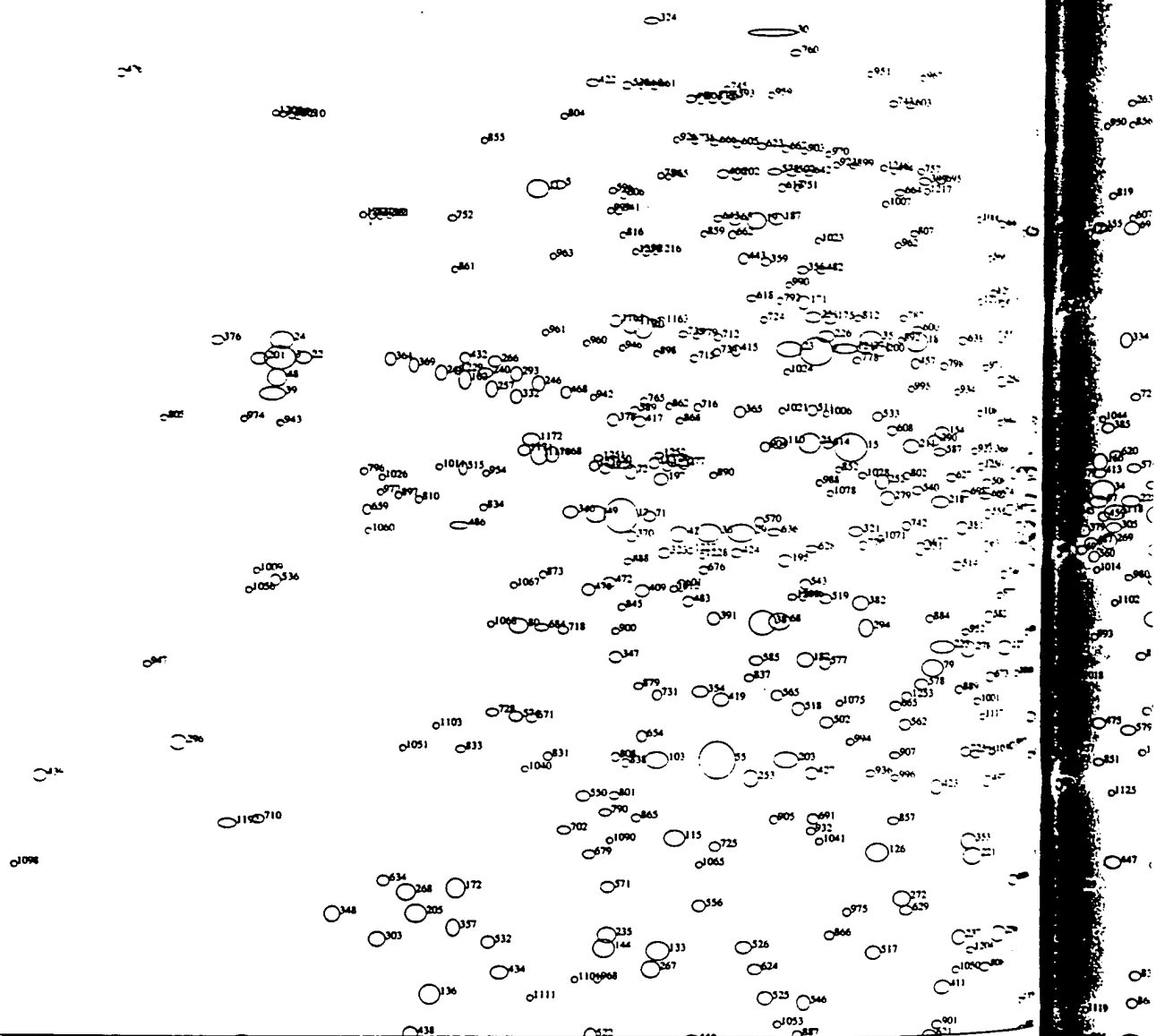


Figure 3. Upper left (high molecular weight, acidic) quadrant (#1) of the rat liver map, showing spot numbers.

24. Up

2

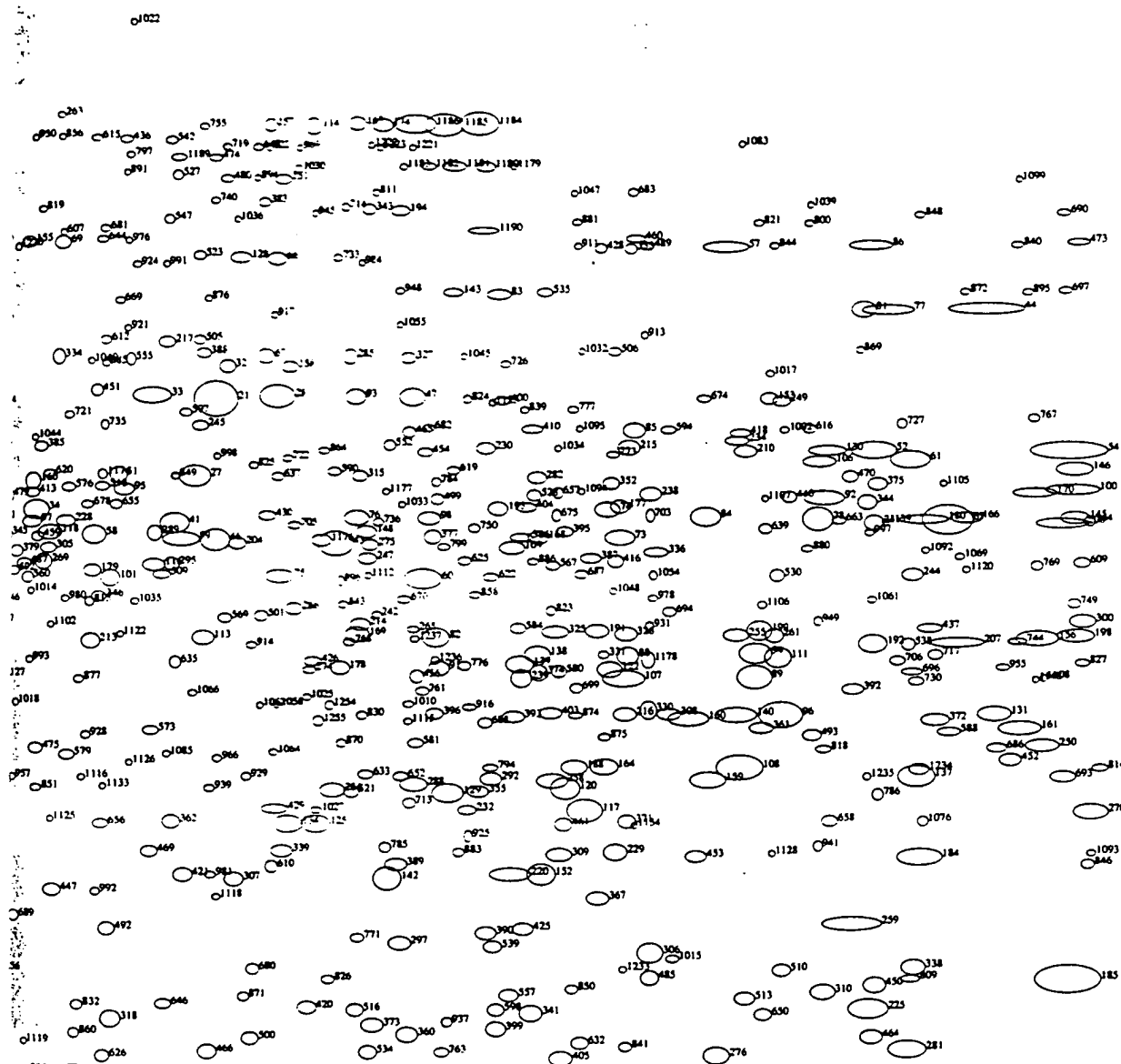


Figure 4. Upper right (high molecular weight, basic) quadrant (#2) of the rat liver map, showing spot numbers.

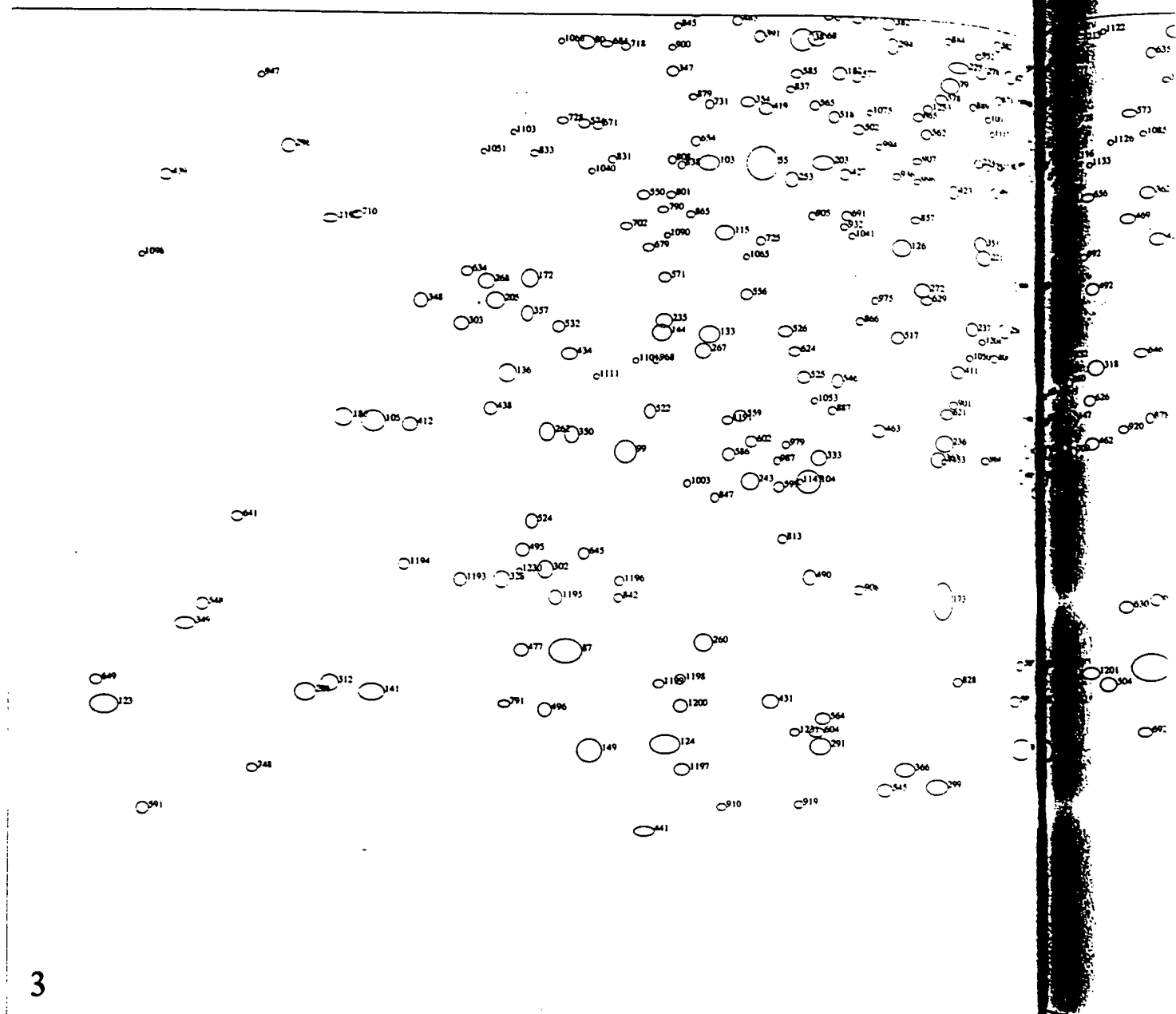


Figure 5. Lower left (low molecular weight, acidic) quadrant (#3) of the rat liver map, showing spot numbers.

Figure 6. Lower r

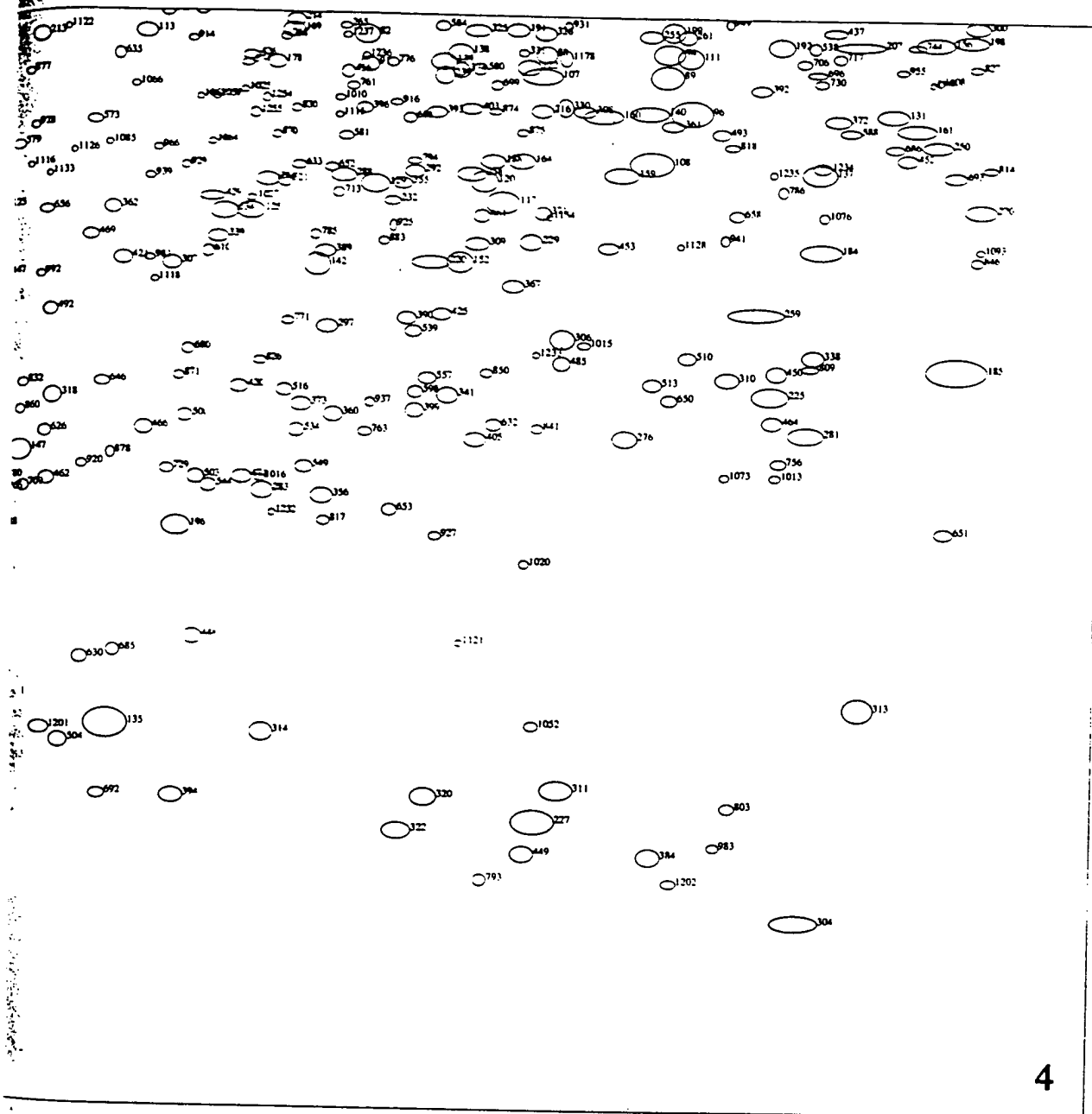


Figure 6. Lower right (low molecular weight, basic) quadrant (#4) of the rat liver map, showing spot numbers.

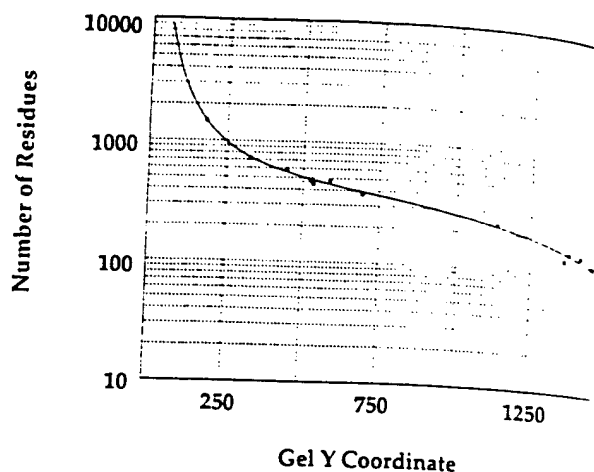
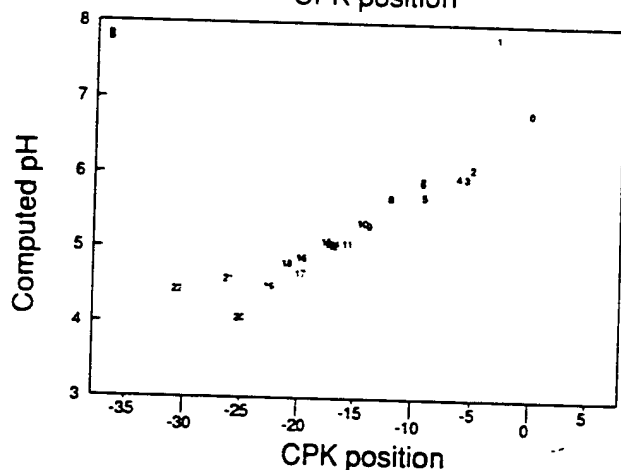
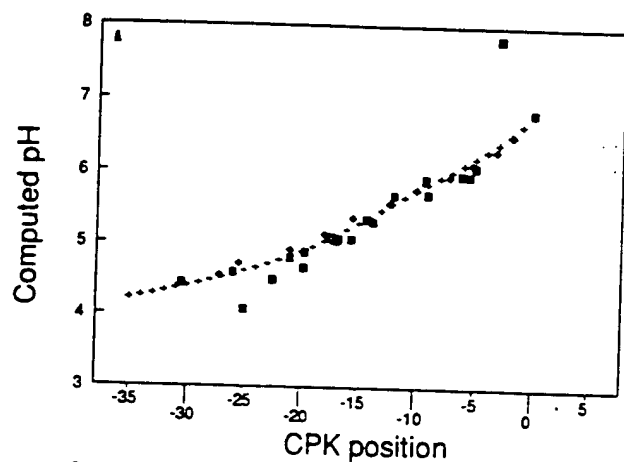


Figure 8. Plot of number of amino acids versus gel Y-position, with fitted curve used to predict molecular mass of unidentified proteins.

Figure 7. (a) Plot of computed isoelectric point versus gel X-position for two sets of carbamylated standard proteins (rabbit muscle CPK [+]) and human hemoglobin  $\beta$  chain, filled diamonds) and several other proteins (shaded squares). (b) The identities of the various proteins represented by the squares are indicated by the numbers in corresponding positions on (a); these refer to Table 4.

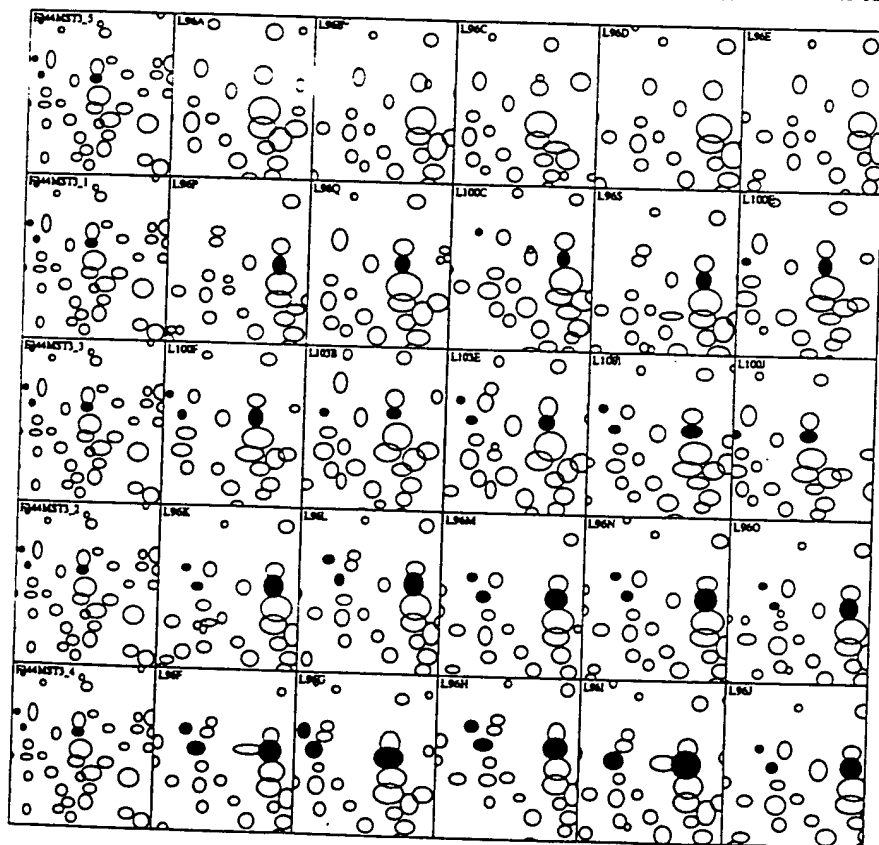


Figure 9. Montage showing effects in the region of MSN:413. The montage shows a small window into one portion of the 2-D pattern, one row of windows for each experimental group, and one panel for each gel in the experiment. The left-most pattern in each row is a group-specific copy of the master pattern followed by the patterns for the five individual rats in the group. The highlighted protein spots (filled circles) are spot 413 (on the right of each panel; identified as cytosolic HMG-CoA synthase) and two modified forms of it (1250 and 933). From the top, the rows (experimental groups) are: high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine.

# Regulation of Rat Liver 413

(Putative Cytosolic HMG-CoA Synthase, 53kd)  
Test Compounds in Diet

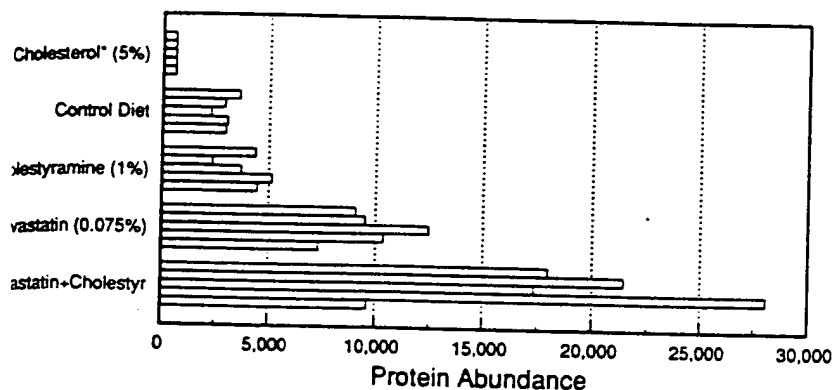


Figure 10. Bargraph showing the quantitative effects of various treatments on the abundance of MSN:413 (cytosolic HMG-CoA synthase) in the gels of Fig. 9.

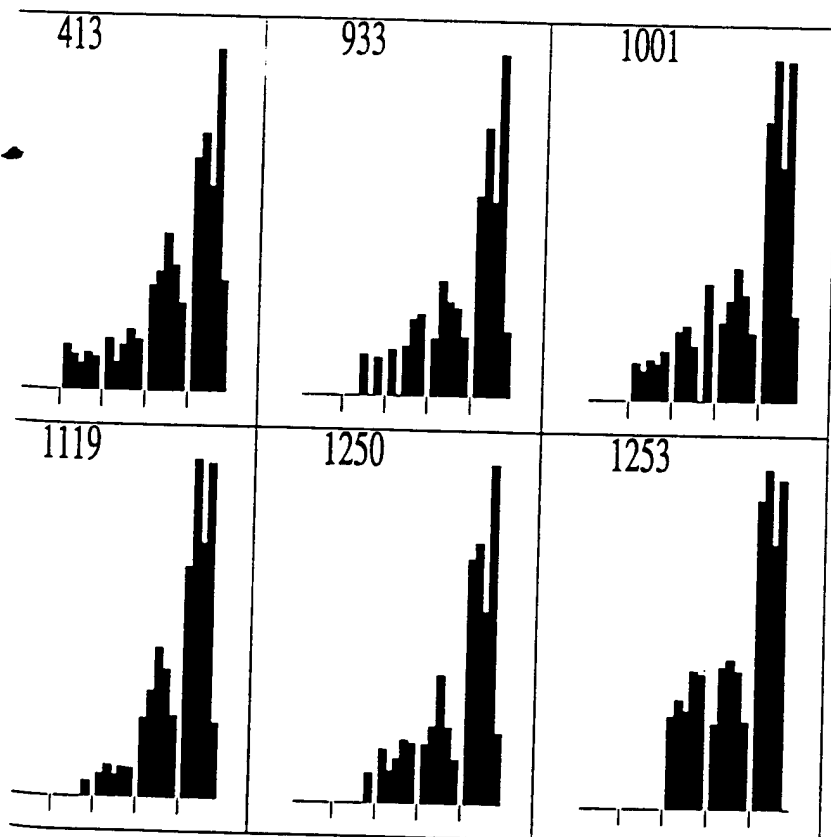


Figure 11. Bargraphs of a series of six coregulated spots including MSN:413. In the bargraphs, the abundances of the appropriate spot (master spot number shown at the top of the panel) in each animal are shown. The five five-animal groups are in the order (left to right): high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine. Each bar within a group represents one experimental animal liver (one 2-D gel). Note the correlated expression of the 6 spots, especially in the two far right (most strongly induced) groups.



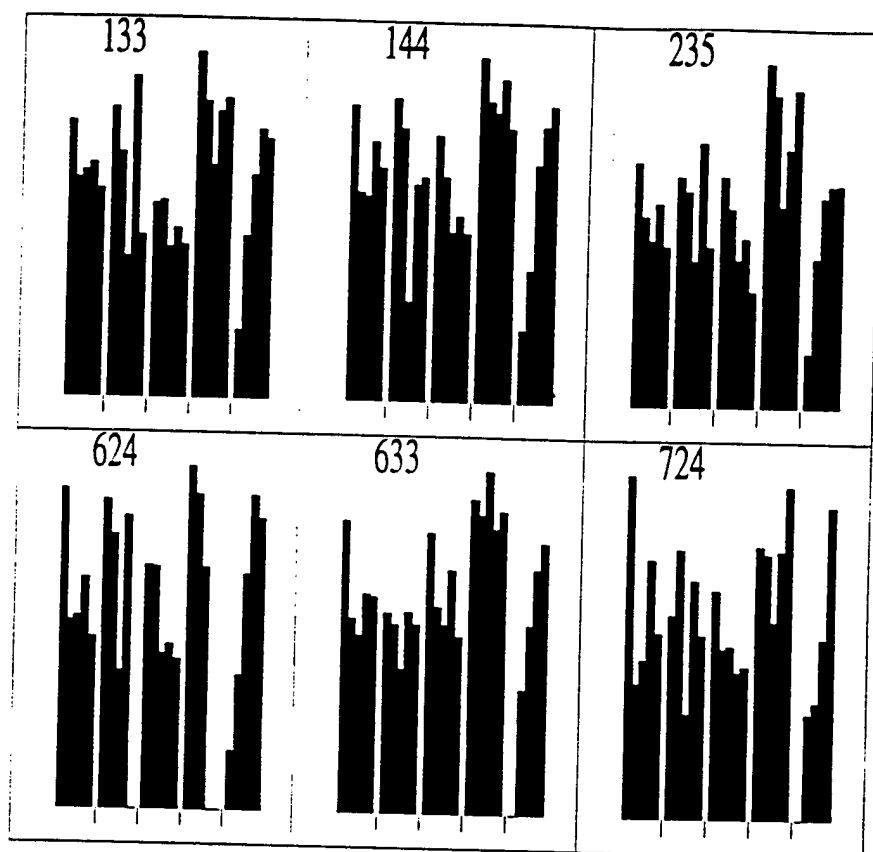


Figure 12. Data on a second coregulated group of spots, presented as in Fig. 11. The fourth experimental group (lovastatin) shows a modest induction, while the fifth group (lovastatin plus cholestyramine) does not.

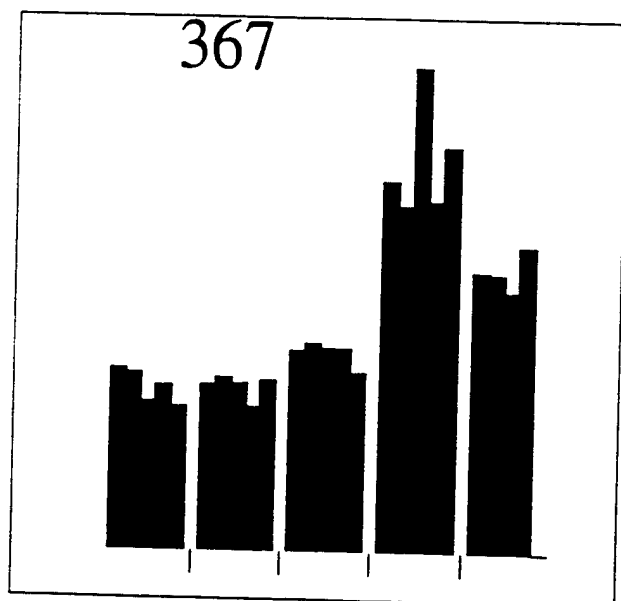


Figure 13. Data on spot MSN:367, presented as in Fig. 11. This protein shows unambiguously the anti-synergistic effect of lovastatin and cholestyramine (fifth group) as compared to lovastatin (fourth group). This response contrasts strongly with the regulation pattern seen in Fig. 11.

Appendix

Mass

x

311
568
812
546
845
626
806
755
646
1204
332
787
313
807
1184
1263
743
768
1215
1145
1037
863
712
763
304
1165
684
1318
1824
1203
139
306
605
621
1113
1820
725
2001
722
678
1682
109
1171
1400
1853
1888
735
1283
1252
779
1064
686
638
1582
1570
1264
1338
1833
1767
925
534
1811
1412
1471
1662
1596
1817
516
1589
1706
651
1415
1773
1338
1708

Inter table of;  
Predicted moie

Table 1. Master table of proteins in the rat liver database<sup>a)</sup>

MSN	X	Y	CPKd	SDSMW	MSN	X	Y	CPKd	SDSMW	MSN	X	Y	CPKd	SDSMW
3	311	434	<-35.0	63,800	95	1119	536	-9.9	53,800	174	1364	183	-6.7	162,900
5	568	263	-24.3	102,900	96	1731	756	-2.0	40,700	175	825	393	-15.7	69,300
8	812	426	-16.0	64,800	97	1033	566	-11.4	51,600	177	1582	553	-3.6	52,600
11	549	268	-25.2	101,000	98	1406	565	-6.1	51,700	178	1321	710	-7.2	43,000
15	845	520	-15.3	55,200	99	578	1149	-23.8	25,000	179	1089	615	-10.4	48,300
17	629	589	-21.6	50,000	100	2004	538	>0.0	53,700	180	1866	567	-0.5	51,600
18	906	414	-14.0	66,300	101	1106	623	-10.1	47,900	181	411	295	-32.1	91,200
19	755	298	-17.5	90,200	102	482	455	-28.5	61,300	182	804	730	-16.2	42,000
20	649	403	-20.9	67,900	103	665	830	-20.2	37,300	184	1860	896	-0.6	34,500
21	1204	448	-8.7	62,100	104	773	1182	-17.0	23,800	185	1997	1017	>0.0	29,800
22	332	434	<-35.0	63,800	105	312	1117	<-35.0	26,100	186	279	1113	<-35.0	26,300
23	787	424	-16.6	65,000	106	1769	509	-1.5	56,100	187	773	296	-17.0	90,800
24	313	417	<-35.0	66,000	107	1585	720	-3.6	42,500	188	1538	807	-4.2	38,400
25	807	516	-16.1	55,500	108	1692	807	-2.4	38,300	191	1560	674	-3.9	44,900
27	1184	524	-8.0	54,900	109	1482	593	-4.8	49,700	192	1818	687	-0.9	44,200
28	1263	446	-8.0	62,400	110	778	516	-16.9	55,500	193	1469	555	-5.0	52,400
29	743	605	-17.8	49,000	111	1728	700	-2.0	43,500	194	1380	266	-6.4	101,600
30	768	112	-17.2	348,600	113	1191	680	-8.9	44,500	195	784	632	-16.7	47,300
32	1216	417	-8.6	66,000	114	1298	185	-7.5	160,800	196	1227	1185	-8.4	23,700
33	1145	445	-9.5	62,500	115	682	907	-19.6	34,100	197	667	553	-20.1	52,600
34	1037	555	-11.3	52,400	116	1146	610	-9.5	48,700	198	2006	681	>0.0	44,500
35	863	412	-14.9	66,600	117	1548	849	-4.1	36,500	199	1711	674	-2.2	44,900
36	712	606	-18.7	48,900	118	1050	577	-11.1	50,800	200	872	424	-14.7	65,000
38	763	694	-17.3	43,800	120	1530	828	-4.3	37,400	201	292	435	<-35.0	63,700
39	304	470	<-35.0	59,800	121	838	423	-15.4	65,200	202	736	253	-18.0	107,800
41	1165	569	-9.2	51,400	122	1572	712	-3.8	42,900	203	786	829	-16.7	37,400
42	684	607	-19.6	48,800	123	23	1433	<-35.0	15,300	204	1224	589	-8.5	50,000
43	1318	589	-7.3	50,000	124	621	1474	-21.9	13,900	205	439	983	-30.9	31,100
44	1924	362	-0.1	74,600	125	1298	862	-7.5	36,000	206	1994	571	>0.0	51,300
46	1203	586	-8.7	50,200	126	872	921	-14.7	33,500	207	1895	687	-0.3	44,200
47	1391	447	-6.3	62,300	127	1000	717	-12.0	42,600	208	240	1418	<-35.0	15,800
48	309	454	<-35.0	61,500	128	1229	311	-8.4	86,100	210	1700	499	-2.3	57,000
49	605	587	-22.5	50,100	129	1422	832	-5.8	37,300	211	902	517	-14.1	55,400
50	621	535	-21.8	53,900	130	1776	499	-1.4	57,000	213	1087	684	-10.4	44,400
51	1113	522	-10.0	55,000	131	1830	757	-0.1	40,700	214	1340	668	-7.0	45,200
52	1820	499	-0.9	57,000	132	660	537	-20.4	53,800	215	1591	495	-3.5	57,300
53	725	177	-18.3	170,800	133	666	1019	-20.2	29,700	216	1585	755	-3.6	40,700
54	2001	500	>0.0	56,900	134	1271	862	-7.9	36,000	217	1159	393	-9.3	69,300
55	722	830	-18.4	37,300	135	1161	1389	-9.3	16,800	218	931	572	-13.5	51,200
56	678	533	-19.8	54,100	136	453	1063	-29.7	28,100	219	713	177	-18.7	170,500
57	1682	302	-2.5	89,000	137	1858	823	-0.6	37,700	220	1479	911	-4.9	33,900
58	1091	580	-10.3	50,600	138	1504	697	-4.6	43,700	221	965	927	-12.8	33,300
59	1171	585	-9.2	50,300	139	1488	707	-4.8	43,200	223	934	716	-13.5	42,700
60	1400	624	-6.2	47,800	140	1689	756	-2.4	40,700	225	1812	1045	-1.0	28,800
61	1853	508	-0.6	56,200	141	311	1417	<-35.0	15,800	226	821	411	-15.8	66,800
62	1888	567	-0.4	51,500	142	1366	915	-6.7	33,800	227	1586	1483	-3.6	13,600
65	735	297	-18.1	90,500	143	1429	346	-5.7	77,900	228	1065	567	-10.8	51,600
66	1263	312	-8.0	85,900	144	615	1017	-22.1	29,800	229	1577	890	-3.7	34,800
67	1252	407	-8.1	67,300	145	2006	566	>0.0	51,600	230	1458	496	-5.2	57,300
68	779	692	-16.8	43,900	146	2006	518	>0.0	55,300	232	1440	849	-5.5	36,500
69	1064	296	-10.8	90,800	147	1070	1108	-10.7	26,500	234	1692	489	-2.4	57,900
71	656	589	-20.6	50,000	148	1347	578	-6.9	50,800	235	618	1004	-22.0	30,300
72	638	545	-21.2	53,100	149	541	1481	-25.7	13,700	236	920	1138	-13.7	25,400
73	1582	583	-3.6	50,400	150	1645	760	-2.8	40,500	237	952	1008	-13.1	30,200
74	1570	556	-3.8	52,300	151	1269	236	-7.9	117,000	238	1611	541	-3.2	53,500
75	1264	621	-8.0	48,000	152	1507	911	-4.5	33,900	239	1489	720	-4.8	42,500
76	1338	564	-7.0	51,800	153	1722	448	-2.1	62,100	240	501	448	-27.7	62,100
77	1833	363	-0.8	74,400	154	932	503	-13.5	56,600	241	1820	569	-0.9	51,400
78	1767	565	-1.5	51,700	155	1031	294	-11.4	91,400	242	1357	658	-6.8	45,800
79	925	738	-13.6	41,600	156	1970	684	>0.0	44,400	243	711	1182	-18.7	23,800
80	534	698	-26.1	43,600	157	1258	183	-8.1	162,400	244	1855	621	-0.6	48,000
81	1811	363	-1.0	74,500	158	1275	417	-7.8	65,900	245	1189	474	-8.9	59,300
82	1412	681	-6.0	44,500	159	1663	820	-2.6	37,800	246	551	459	-25.1	61,000
83	1471	347	-5.0	77,500	160	1034	527	-11.4	54,600	247	1348	604	-6.9	49,100
84	1662	563	-2.7	51,800	161	1953	771	>0.0	40,000	248	460	448	-29.3	62,100
85	1596	479	-3.4	58,900	162	1020	1482	-11.6	13,700	249	1733	451	-1.9	61,800
86	1817	301	-0.9	89,100	164	1566	806	-3.8	38,400	250	1974	788	>0.0	39,200
87	516	1371	-27.0	17,400	166	1905	565	-0.2	51,700	251	808	392	-16.1	69,500
88	1589	698	-3.5	43,600	167	1340	181	-7.0	164,900	252	874	553	-14.6	52,500
89	1706	719	-2.2	42,500	168	1506	583	-4.6	50,400	253	753	848	-17.6	36,500
90	651	329	-20.8	81,700	169	1338	678	-7.0	44,700	254	995	450	-12.1	61,900
91	1415	710	-6.0	43,000	170	1969	541	>0.0	53,500	255	1690	679	-2.4	44,600
92	1773	545	-1.4	53,200	171	800	378	-16.3	71,800	256	994	1006	-12.1	30,200
93	1338	446	-7.0	62,300	172	476	958	-28.7	32,100	257	508	464	-27.4	60,400
94	1708	696	-2.2	43,700	173	919	1314	-13.7	19,300	258	1517	820	-4.4	37,800

Master table of proteins in the rat liver database, showing spot master number, gel position (x and y), isoelectric point relative to CPK standards, and predicted molecular mass (from the standard curve of Fig. 8).

MSN	X	Y	CPKpl	SDSMW	MSN	X	Y	CPKpl	SDSMW	MSN	X	Y	CPKpl	SDSMW	MSN	X
259	1796	961	-1.1	31,900	345	1006	578	-11.9	50,800	426	1296	704	-7.6	43,300	809	
260	661	1361	-20.4	17,700	346	1095	640	-10.3	46,800	427	810	843	-16.0	36,800	1099	
261	1725	679	-2.0	44,600	347	625	728	-21.7	42,000	428	1565	303	-3.9	88,700	1696	
262	496	1127	-28.0	25,800	348	361	983	-35.3	31,100	429	1259	847	-8.0	36,800	948	
263	1063	172	-10.9	177,400	349	110	1343	<-35.0	18,300	430	1253	562	-8.1	51,900	481	
265	1390	673	-6.3	45,000	350	521	1130	-26.7	25,700	431	734	1426	-18.1	15,500	1334	
266	510	437	-27.3	63,400	351	912	619	-13.9	48,100	432	483	433	-28.5	63,900	868	
267	660	1038	-20.4	29,000	352	1574	530	-3.7	54,300	434	518	1041	-26.9	28,900	798	
268	430	961	-31.0	31,900	353	961	912	-12.9	33,900	435	1020	1170	-11.6	24,300	822	
269	1044	606	-11.2	48,900	354	706	762	-18.9	40,400	436	1122	196	-9.8	147,600	632	
270	2019	853	>0.0	36,300	355	1450	830	-5.3	37,300	437	1870	673	-0.5	45,000	1332	
271	857	422	-15.0	65,200	356	1374	1152	-6.5	24,900	438	435	1102	-31.0	26,700	603	
272	895	968	-14.2	31,700	357	474	997	-28.7	30,600	439	86	847	<-35.0	36,600	1190	
274	1292	712	-7.6	42,900	358	798	346	-16.3	77,800	440	1740	544	-1.8	53,200	479	
275	1350	590	-6.9	49,900	359	764	338	-17.3	79,400	441	599	1571	-22.8	10,800	768	
276	1670	1069	-2.6	27,100	360	1384	1068	-6.4	27,900	443	743	335	-17.8	80,100	747	
277	688	538	-19.4	53,700	361	1713	769	-2.1	40,100	446	801	668	-16.2	45,200	1170	
278	961	718	-13.0	42,600	362	1161	859	-9.3	36,100	447	1050	926	-11.1	33,300	1502	
279	879	570	-14.5	51,300	363	914	1156	-13.8	24,800	448	1245	1298	-8.2	19,800	1728	
281	1848	1084	-0.7	27,300	364	412	435	-32.0	63,700	449	1576	1516	-3.7	12,600	507	
282	1505	525	-4.6	54,800	365	741	486	-17.9	58,200	450	1818	1021	-0.9	29,600	870	
283	1313	1147	-7.3	25,100	366	878	1503	-14.6	13,000	451	1094	440	-10.3	63,100	1347	
284	1314	829	-7.3	37,400	367	1560	935	-3.9	33,000	452	1945	802	>0.0	38,600	1513	
285	1332	408	-7.1	67,200	368	983	520	-12.4	55,200	453	1652	894	-2.8	34,600	308	
286	1277	652	-7.8	46,100	369	434	441	-31.0	63,000	454	1403	500	-6.1	56,900	1851	
288	1391	824	-6.3	37,600	370	639	610	-21.2	48,700	456	1394	718	-6.3	42,600	1463	
289	1147	579	-9.5	50,700	371	1587	860	-3.6	36,100	457	905	436	-14.0	63,500	909	
290	925	511	-13.6	55,900	372	1875	762	-0.5	40,400	459	1038	581	-11.3	50,500	625	
291	787	1476	-16.6	13,900	373	1351	1059	-6.8	28,300	460	1598	294	-3.4	91,400	1164	
292	1462	818	-5.1	37,800	374	1506	715	-4.6	42,700	461	1528	863	-4.3	35,900	803	
293	531	449	-26.3	62,000	375	1823	532	-0.9	54,200	462	1098	1137	-10.2	25,400	1259	
294	860	698	-14.9	43,600	376	254	417	<-35.0	65,900	463	849	1125	-15.2	25,800	856	
295	1162	609	-9.3	48,700	377	1409	583	-6.1	50,400	464	1814	1072	-0.9	27,800	803	
296	218	814	<-35.0	38,000	378	621	494	-21.8	57,500	465	1388	481	-6.3	58,700	1162	
297	1377	979	-6.5	31,300	379	1017	595	-11.7	49,600	466	1194	1084	-8.9	27,300	128	
299	913	1523	-13.9	12,400	381	953	598	-13.1	49,400	468	577	467	-23.9	60,100	1355	
300	2012	667	>0.0	45,300	382	856	674	-15.0	44,900	469	1140	888	-9.6	34,900	595	
301	702	178	-19.0	169,200	383	1252	258	-8.1	105,300	470	1797	524	-1.1	54,800	1369	
302	494	1280	-28.1	20,400	384	1699	1518	-2.3	12,500	471	1293	1133	-7.6	25,500	992	
303	403	1008	-32.6	30,100	385	1042	493	-11.2	57,500	472	618	655	-21.9	46,000	1125	
304	1843	1585	-0.7	10,300	386	1490	583	-4.7	50,400	473	2009	299	>0.0	89,900	705	
305	1049	593	-11.1	49,800	387	1554	603	-4.0	49,100	474	1205	215	-8.7	131,300	1477	
306	1608	989	-3.3	30,900	388	1193	404	-8.9	67,700	475	1035	788	-11.4	39,200	980	
307	1219	916	-8.5	33,700	389	1374	902	-6.5	34,300	476	160	155	<-35.0	207,600	700	
308	1627	755	-3.0	40,700	390	1456	969	-5.2	31,700	477	469	1370	-28.9	17,400	1028	
309	1524	892	-4.4	34,700	391	718	690	-18.5	44,000	478	599	662	-22.8	45,600	898	
310	1769	1028	-1.5	29,400	392	1799	732	-1.1	41,900	479	1009	540	-11.8	53,500	789	
311	1609	1451	-3.3	14,700	393	1482	758	-4.8	40,600	480	1216	235	-8.6	117,400	777	
312	266	1408	<-35.0	16,100	394	1227	1461	-8.4	14,400	482	816	346	-15.9	77,800	980	
313	1902	1365	-0.3	17,600	395	1530	577	-4.3	50,800	483	683	673	-19.3	44,900	1519	
314	1316	1395	-7.3	16,600	396	1410	755	-6.0	40,800	485	1608	1013	-3.3	30,000	1212	
315	1341	523	-7.0	54,900	397	912	256	-13.9	106,400	486	478	599	-28.6	49,300	760	
318	1104	1053	-10.1	28,500	399	1465	1063	-5.0	28,100	487	1025	607	-11.5	48,800	618	
320	1480	1459	-4.9	14,400	400	1473	450	-4.9	61,900	488	1045	1186	-11.2	23,700	1142	
321	850	603	-15.1	49,100	401	1029	1140	-11.5	25,300	489	1609	301	-3.3	89,200	532	
322	1454	1484	-5.3	13,300	403	1516	754	-4.4	40,800	490	775	1289	-17.0	20,100	771	
323	670	626	-20.0	47,700	404	1495	554	-4.7	52,500	491	692	178	-19.3	169,300	1068	
324	655	101	-20.6	420,500	405	1525	1092	-4.3	27,100	492	1100	964	-10.2	31,800	822	
325	1521	675	-4.4	44,800	406	723	252	-18.4	108,000	493	1760	776	-1.6	39,700	914	
326	1587	677	-3.6	44,700	409	650	663	-20.8	45,500	494	882	247	-14.5	110,700	1064	
327	1388	408	-6.3	67,000	410	1501	478	-4.6	59,000	495	470	1258	-28.9	21,200	1524	
328	448	1291	-30.0	20,100	411	936	1057	-13.4	28,300	496	494	1436	-28.1	15,200	1392	
330	1608	751	-3.3	40,900	412	350	1120	-35.9	26,000	497	980	852	-12.5	36,400	982	
331	1566	697	-3.8	43,700	413	1033	538	-11.4	53,700	499	1414	546	-6.0	53,100	1487	
332	531	471	-26.3	59,600	415	737	425	-18.0	64,900	500	1234	1072	-8.3	27,800	758	
333	784	1156	-16.7	24,700	416	1578	606	-3.7	48,900	501	1246	659	-8.2	45,700	687	
334	1059	407	-10.9	67,300	417	646	496	-21.0	57,300	502	824	792	-15.7	39,000	930	
335	1593	303	-3.5	88,500	418	1695	482	-2.3	58,600	503	1246	1134	-8.2	25,500	1888	
336	1616	598	-3.2	49,400	419	725	770	-18.3	40,000	504	1115	1407	-9.9	16,200	642	
338	1854	1004	-0.6	30,300	420	1289	1041	-7.7	28,900	505	1189	391	-8.9	69,700	1317	
339	1265	888	-8.0	34,900	421	1171	912	-9.1	33,900	506	1578	402	-3.7	68,000	65	
340	581	585	-23.6	50,300	422	599	162	-22.8	193,700	507	787	250	-16.6	108,000	1014	
341	1497	1047	-4.7	28,700	423	929	856	-13.6	36,200	508	979	552	-12.5	52,600	732	
343	1351	265	-6.8	102,200	424	739	625	-17.9	47,700	509	1153	619	-9.4	48,100	1627	
344	1813	549	-0.9	52,800	425	1490	965	-4.7	31,800	510	1730	1006	-2.0	30,200	1009	

MSN	X	Y	CPKd	SDSMW
511	809	484	-16.0	58,400
512	1099	533	-10.2	54,100
513	1696	1034	-2.3	29,200
514	948	636	-13.2	47,100
515	481	543	-28.5	53,400
516	1334	1044	-7.1	28,800
517	868	1021	-14.8	29,700
518	798	779	-16.3	39,600
519	822	670	-15.7	45,100
520	632	165	-21.5	189,000
521	1332	830	-7.1	37,300
522	603	1104	-22.6	26,600
523	1190	309	-8.9	86,800
524	479	1226	-28.6	22,300
525	768	1066	-17.2	28,000
526	747	1016	-17.7	29,800
527	1170	231	-9.2	119,600
528	1502	542	-4.6	53,400
530	1728	620	-2.0	48,000
532	507	1011	-27.4	30,000
533	870	489	-14.7	57,900
534	1347	1085	-6.9	27,300
535	1513	346	-4.5	77,800
536	308	654	<-35.0	46,000
538	1851	689	-0.7	44,100
539	1463	982	-5.1	31,100
540	909	561	-13.9	52,000
541	625	289	-21.7	93,100
542	1164	198	-9.2	146,200
543	803	655	-16.2	45,900
544	1259	1143	-8.0	25,200
545	856	1526	-15.0	12,200
546	803	1071	-16.2	27,800
547	1162	274	-9.3	98,400
548	128	1321	<-35.0	19,000
549	1355	1122	-6.8	25,900
550	595	866	-23.0	35,800
552	1369	494	-6.6	57,500
553	992	405	-12.2	67,600
555	1125	410	-9.8	66,900
556	705	975	-18.9	31,400
557	1477	1030	-4.9	29,300
558	980	583	-12.5	50,400
559	700	1109	-19.1	26,400
560	1028	621	-11.5	48,000
562	898	794	-14.1	38,900
564	789	1446	-16.6	14,900
565	777	766	-16.9	40,200
566	980	328	-12.5	81,900
567	1519	611	-4.4	48,600
569	1212	661	-8.6	45,600
570	760	594	-17.4	49,700
571	618	956	-21.9	32,100
573	1142	771	-9.6	40,000
574	532	787	-26.2	39,300
575	771	250	-17.1	109,200
576	1068	534	-10.8	54,100
577	822	734	-15.7	41,800
578	914	754	-13.8	40,800
579	1064	794	-10.8	38,900
580	1524	714	-4.4	42,800
581	1392	783	-6.3	39,400
582	982	686	-12.4	44,200
584	1487	672	-4.8	45,000
585	758	731	-17.4	41,900
586	687	1152	-19.5	24,900
587	930	523	-13.5	55,000
588	1888	774	-0.4	39,900
589	642	485	-21.1	58,300
590	1317	519	-7.3	55,300
591	65	1548	<-35.0	11,500
592	1014	614	-11.7	48,400
593	732	176	-18.1	172,300
594	1627	478	-3.0	59,000
595	1009	1426	-11.8	15,500

MSN	X	Y	CPKd	SDSMW
596	619	269	-21.9	100,500
597	1176	461	-9.1	60,700
598	1465	1044	-5.0	28,800
599	741	1188	-17.9	23,600
600	907	402	-14.0	68,000
601	687	658	-19.5	45,800
602	712	1138	-18.7	25,400
603	898	181	-14.1	165,200
604	783	1461	-16.7	14,400
605	736	223	-18.0	125,300
606	629	273	-21.6	98,700
607	1064	286	-10.8	94,000
608	883	503	-14.5	56,700
609	2012	610	>0.0	48,700
610	1255	903	-8.1	34,200
612	1103	391	-10.1	69,600
613	778	265	-16.9	102,000
614	824	518	-15.7	55,400
615	1095	195	-10.3	149,100
616	1759	478	-1.6	59,000
617	994	372	-12.1	72,900
618	751	374	-17.6	72,400
619	1429	518	-5.7	55,300
620	1050	520	-11.1	55,200
621	923	1105	-13.7	26,600
622	1462	622	-5.1	47,900
623	759	225	-17.4	124,000
624	758	1038	-17.4	29,000
625	1438	606	-5.5	48,900
626	1096	1089	-10.2	27,200
627	942	548	-13.3	53,000
628	809	621	-16.0	48,000
629	899	979	-14.1	31,300
630	1135	1321	-9.6	19,100
631	979	615	-12.5	48,300
632	1542	1076	-4.1	27,600
633	1345	814	-6.9	38,000
634	409	950	-32.2	32,400
635	1165	704	-9.2	43,300
636	774	604	-17.0	49,000
637	1263	524	-8.0	54,800
638	952	411	-13.1	66,700
639	1717	575	-2.1	51,000
640	994	292	-12.1	92,000
641	165	1224	<-35.0	22,400
642	803	251	-16.2	108,900
643	719	296	-18.5	90,700
644	1100	294	-10.2	91,400
645	534	1263	-26.1	21,000
646	1153	1038	-9.4	29,000
648	1246	204	-8.2	140,000
649	14	1406	<-35.0	16,200
650	1713	1049	-2.1	28,600
651	1986	1183	>0.0	23,800
652	1378	816	-6.5	38,000
653	1442	1165	-5.5	24,400
654	650	806	-20.8	38,400
655	1111	551	-10.0	52,700
656	1095	861	-10.3	36,000
657	1524	540	-4.4	53,600
658	1777	860	-1.4	36,000
659	391	584	-33.4	50,400
660	977	565	-12.5	51,700
661	658	166	-20.5	187,500
662	732	312	-18.1	86,100
663	1787	567	-1.2	51,500
664	888	268	-14.4	100,900
665	889	775	-14.3	39,800
666	715	221	-18.6	126,300
667	781	227	-16.8	122,400
668	646	165	-21.0	189,100
669	1116	353	-9.9	76,300
670	1382	643	-6.4	46,600
671	547	789	-25.3	39,200
673	984	746	-12.4	41,200

MSN	X	Y	CPKd	SDSMW
674	1661	448	-2.7	62,100
675	1523	562	-4.4	51,900
676	708	642	-18.8	46,700
677	919	615	-13.7	48,300
678	1085	551	-10.5	52,700
679	600	923	-22.7	33,400
680	1237	1004	-8.3	30,300
681	1103	283	-10.1	95,100
682	1406	477	-6.1	59,100
683	1596	249	-3.4	109,800
684	555	699	-24.8	43,500
685	1167	1313	-9.2	19,300
686	1932	790	0.0	39,100
687	1545	619	-4.1	48,100
688	1456	764	-5.2	40,300
689	1011	953	-11.8	32,300
690	1995	270	>0.0	100,200
691	812	888	-16.0	34,900
692	1154	1461	-9.4	14,400
693	1993	819	>0.0	37,800
694	1628	656	-3.0	45,900
695	928	254	-13.6	107,000
696	1854	715	-0.6	42,700
697	1997	345	>0.0	78,000
698	957	563	-13.0	51,800
699	1540	730	-4.2	42,000
702	577	900	-23.8	34,400
703	1610	562	-3.2	51,900
705	1278	571	-7.8	51,200
706	1841	704	-0.7	43,300
707	1018	1386	-11.7	16,900
709	1074	1145	-10.7	25,100
710	293	889	<-35.0	34,800
712	720	412	-18.5	66,600
713	1386	841	-6.4	36,800
714	1328	263	-7.1	103,100
715	698	433	-19.1	63,900
716	701	481	-19.0	58,700
717	1875	699	-0.5	43,600
718	575	702	-23.9	43,400
719	1216	204	-8.6	140,400
721	1069	464	-10.8	60,400
722	1272	506	-7.9	56,400
723	958	822	-13.0	37,700
724	763	395	-17.3	69,100
725	720	916	-18.5	33,700
726	1476	415	-4.9	66,200
727	1846	473	-0.7	59,400
728	510	783	-27.3	39,400
729	1217	1126	-8.6	25,800
730	1858	724	-0.6	42,300
731	665	765	-20.2	40,300
733	1321	312	-7.2	85,900
734	719	427	-18.5	64,600
735	1101	473	-10.2	59,500
736	1359	569	-6.7	51,400
738	696	220	-19.2	127,600
739	687	409	-19.5	67,000
740	1205	256	-8.7	106,200
741	995	563	-12.1	51,900
742	898	596	-14.1	49,500
743	881	181	-14.5	165,900
744	1951	686	>0.0	44,200
745	726	168	-18.3	183,600
746	999	643	-12.0	46,600
748	182	1503	<-35.0	13,000
749	2005	649	>0.0	46,300
750	1448	575	-5.4	51,000
751	792	266	-16.5	101,800
752	469	296	-28.9	90,600
754	664	254	-20.3	107,000
755	1195	184	-8.8	161,000
756	1821	1113	-0.9	26,300
757	909	246	-13.9	111,000
760	790	133	-16.5	264,900

MSN	X	Y	CPKpl	SDSMW	MSN	X	Y	CPKpl	SDSMW	MSN	X	Y	CPKpl	SDSMW	MSN	X
761	1399	733	-6.2	41,800	848	1863	271	-0.6	99,500	939	1197	827	-8.8	37,500	940	405
763	1416	1085	-5.9	27,300	849	1166	523	-9.2	54,900	941	1765	885	-1.5	35,000	941	1298
764	2020	569	>0.0	51,400	850	1535	1024	-4.2	29,600	942	602	472	-22.7	59,600	942	856
765	651	475	-20.8	59,300	851	1035	826	-11.4	37,500	943	312	498	<-35.0	57,100	943	1284
766	1052	1149	-11.1	25,000	852	834	542	-15.5	53,400	944	993	491	-12.1	57,700	944	986
767	1968	468	>0.0	59,900	855	499	220	-27.8	127,100	945	1300	269	-7.5	100,300	945	1547
768	1330	685	-7.1	44,300	856	1063	194	-10.9	150,500	946	630	423	-21.6	65,100	946	1381
769	1970	613	>0.0	48,500	857	887	890	-14.4	34,800	947	187	736	<-35.0	41,600	947	1525
770	857	617	-15.0	48,200	858	1448	639	-5.4	46,900	948	1380	344	-6.5	78,200	948	1128
771	1337	974	-7.0	31,500	859	706	311	-18.9	86,200	949	1766	665	-1.5	45,400	949	1226
773	1576	502	-3.7	56,700	860	1070	1066	-10.7	28,000	950	1038	193	-11.3	151,000	950	1761
775	969	824	-12.8	37,600	861	472	347	-28.8	77,600	951	860	152	-14.9	213,000	951	541
776	1438	708	-5.5	43,100	862	674	480	-19.9	58,800	952	957	701	-13.0	43,400	952	818
777	1539	458	-4.2	61,000	864	1307	499	-7.4	57,000	954	503	547	-27.6	53,000	954	1036
778	850	434	-15.1	63,800	865	645	887	-21.0	34,900	955	1938	712	>0.0	42,900	955	1439
779	700	411	-19.1	66,800	866	827	1004	-15.6	30,300	957	1010	816	-11.8	37,900	957	1540
780	1052	1136	-11.1	25,500	868	685	494	-19.5	57,400	959	768	174	-17.2	174,900	959	1576
784	1413	529	-6.0	54,400	869	1807	402	-1.0	68,000	960	596	419	-23.0	65,700	960	1089
785	1364	885	-6.7	35,000	870	1323	783	-7.2	39,400	961	557	409	-24.8	67,100	961	949
786	1822	835	-0.9	37,100	871	1228	1031	-8.4	29,300	962	887	320	-14.4	83,900	962	426
787	893	392	-14.3	69,500	872	1904	346	-0.3	77,700	963	564	334	-24.5	80,500	963	1583
790	616	882	-22.0	35,100	873	556	647	-24.8	46,400	964	969	1155	-12.8	24,800	964	779
791	451	1429	-29.8	15,400	874	1540	756	-4.2	40,700	965	671	255	-20.0	106,600	965	1613
792	777	377	-16.9	72,000	875	1566	777	-3.8	39,700	966	1204	798	-8.7	38,700	966	1380
793	1536	1543	-4.2	11,700	876	1198	351	-8.8	76,800	967	910	154	-13.9	210,300	967	284
794	1461	807	-5.1	38,300	877	1076	720	-10.6	42,500	968	609	1048	-22.3	28,700	968	1261
796	388	546	-33.6	53,100	878	1161	1111	-9.3	26,400	969	1285	206	-7.7	138,900	969	393
797	1126	212	-9.8	133,700	879	647	757	-20.9	40,700	970	822	232	-15.8	119,300	970	1817
798	933	437	-13.5	63,400	880	1756	594	-1.6	49,700	971	976	437	-12.6	63,400	971	1245
799	1420	593	-5.9	49,800	881	1543	278	-4.1	97,100	972	403	567	-32.6	51,600	972	1258
800	1759	279	-1.6	96,500	883	1432	890	-5.7	34,800	974	279	495	<-35.0	57,400	974	705
801	624	865	-21.7	35,800	884	922	689	-13.7	44,100	975	844	981	-15.3	31,200	975	1181
802	898	547	-14.2	53,000	885	1103	414	-10.1	66,400	976	1124	295	-9.8	91,100	976	529
803	1775	1468	-1.4	14,200	886	1501	607	-4.6	48,900	977	994	664	-12.1	45,400	977	508
804	573	196	-24.0	148,400	887	798	1103	-16.3	26,600	978	1612	642	-3.2	46,700	978	1898
805	203	494	<-35.0	57,400	888	636	634	-21.3	47,200	979	749	1141	-17.7	25,300	979	873
806	980	1039	-12.5	29,000	889	951	759	-13.1	40,600	980	1064	642	-10.8	46,700	980	1768
807	902	308	-14.1	87,200	890	717	548	-18.6	52,900	981	1197	911	-8.8	33,900	981	836
808	625	827	-21.7	37,500	891	1123	229	-9.8	121,200	983	1762	1508	-1.6	12,800	983	1863
809	1851	1015	-0.7	29,900	892	891	413	-14.3	66,400	984	1344	317	-6.9	84,700	984	826
810	440	573	-30.9	51,100	894	1245	234	-8.2	117,800	985	1024	1105	-11.5	26,600	985	971
811	1358	249	-6.8	109,700	895	1962	346	>0.0	77,700	987	739	1159	-17.9	24,600	987	1697
812	851	393	-15.1	69,400	896	1322	626	-7.2	47,700	988	816	555	-15.9	52,400	988	1157
813	745	1246	-17.8	21,600	897	420	570	-31.4	51,300	990	785	361	-16.7	74,900	990	620
814	2028	810	>0.0	38,200	898	662	428	-20.3	64,500	991	1159	317	-9.3	84,500	991	1867
815	1066	645	-10.4	46,500	899	845	243	-15.3	113,000	992	1090	928	-10.4	33,300	992	2019
816	629	313	-21.6	85,700	900	624	703	-21.7	43,400	993	1030	701	-11.5	43,400	993	1546
817	1376	1177	-6.5	24,000	901	931	1094	-13.5	27,000	994	847	811	-15.2	38,200	994	1545
818	1771	790	-1.4	39,100	903	799	229	-16.3	121,000	995	902	461	-14.1	60,700	995	61
819	1045	263	-11.2	103,100	904	765	520	-17.2	55,200	996	888	847	-14.4	36,600	996	1954
820	984	362	-12.4	74,600	905	775	889	-17.0	34,800	997	1815	579	-0.9	50,700	997	588
821	1712	279	-2.2	96,700	907	888	824	-14.4	37,600	998	1205	504	-8.7	56,500	998	1050
822	1256	205	-8.1	139,200	908	828	1303	-15.6	19,700	999	617	289	-22.0	93,100	999	457
823	1517	654	-4.4	46,000	910	681	1544	-19.7	11,700	1000	968	290	-12.8	92,700	1000	1884
824	1442	449	-5.5	62,000	911	1544	301	-4.1	89,100	1001	970	771	-12.7	40,000	1001	1714
825	1240	513	-8.3	55,800	913	1606	387	-3.3	70,400	1002	1736	478	-1.9	58,900	1002	1717
826	1309	1014	-7.4	29,900	914	1237	688	-8.3	44,100	1003	643	1184	-21.1	23,700	1003	1976
827	2012	708	>0.0	43,100	916	1442	749	-5.5	41,100	1006	822	487	-15.8	58,100	1006	547
828	937	1405	-13.4	16,200	917	1260	367	-8.0	73,700	1007	875	279	-14.6	96,400	1007	1348
830	1342	756	-7.0	40,700	919	764	1541	-17.3	11,700	1009	291	644	<-35.0	46,600	1009	1385
831	562	826	-24.5	37,500	920	1133	1123	-9.7	25,900	1010	1386	745	-6.4	41,200	1010	1078
832	1073	1039	-10.7	29,000	921	1123	380	-9.8	71,500	1011	459	541	-29.4	53,500	1011	975
833	481	820	-28.5	37,800	923	829	242	-15.6	113,200	1012	679	661	-19.7	45,600	1012	1202
834	501	581	-27.8	50,500	924	1131	318	-9.7	84,300	1013	1818	1128	-0.9	25,800	1013	1022
837	751	748	-17.6	41,100	925	1441	874	-5.5	35,400	1014	1032	634	-11.4	47,200	1014	1905
838	635	833	-21.3	37,200	926	679	219	-19.7	128,200	1015	1629	994	-3.0	30,700	1015	1512
839	1494	459	-4.7	60,900	927	1487	1191	-4.8	23,500	1016	1311	1134	-7.4	25,500	1016	1114
840	1952	301	>0.0	89,300	928	1082	775	-10.5	39,800	1017	1722	424	-2.0	65,000	1017	1464
841	1585	1080	-3.6	27,500	929	1231	816	-8.4	38,000	1018	1015	743	-11.7	41,300	1018	1048
842	571	1312	-24.1	19,400	931	1609	670	-3.3	45,100	1020	1574	1219	-3.7	22,500	1020	1122
843	1325	649	-7.2	46,300	932	810	900	-16.0	34,400	1021	781	484	-16.8	58,400	1021	1722
844	1727	301	-2.0	89,200	933	965	520	-12.8	55,100	1022	1129	83	-9.7	591,300	1022	1098
845	630	679	-21.5	44,600	934	947	462	-13.2	60,600	1023	812	317	-15.9	84,600	1023	1830
846	2016	905	>0.0	34,200	936	865	843	-14.8	36,800	1024	785	446	-16.7	62,400	1024	764
847	673	1200	-19.9	23,200	937	1421	1056	-5.9	28,400	1025	1290	739	-7.7	41,500	1025	1968

MSN	X	Y	CPKd	SDSMW	MSN	X	Y	CPKd	SDSMW	MSN	X	Y	CPKd	SDSMW
1028	405	552	-32.3	52,600	1153	921	1158	-13.7	24,700	1246	547	577	-25.3	50,800
1027	1298	848	-7.5	36,500	1154	1594	864	-3.5	35,900	1247	530	576	-26.3	50,900
1028	856	547	-15.0	53,000	1161	637	400	-21.3	68,400	1249	516	572	-27.0	51,200
1030	1284	226	-7.7	123,200	1162	623	397	-21.8	68,800	1250	973	536	-12.7	53,900
1031	986	822	-12.3	37,700	1163	665	397	-20.2	68,700	1251	607	532	-22.4	54,200
1032	1547	403	-4.1	67,900	1168	564	528	-24.4	54,500	1252	665	529	-20.2	54,400
1033	1381	551	-6.4	52,700	1170	552	529	-25.0	54,500	1253	899	766	-14.1	40,200
1034	1525	496	-4.3	57,200	1171	538	524	-25.9	54,800	1254	1311	746	-7.4	41,200
1035	1128	645	-9.7	46,500	1172	545	514	-25.5	55,700	1255	1300	761	-7.5	40,400
1036	1226	274	-8.5	98,300	1174	1099	522	-10.2	55,000	1257	1938	712	0.0	42,900
1039	1761	262	-1.6	103,600	1176	1304	586	-7.5	50,200	1258	1806	718	-1.0	42,600
1040	541	839	-25.7	36,900	1177	1366	539	-6.6	53,700	1259	1727	715	-2.0	42,700
1041	818	910	-15.8	34,000	1178	1608	702	-3.3	43,400	1260	1629	713	-3.0	42,800
1044	1036	485	-11.3	58,300	1179	1485	224	-4.8	124,900	1261	1555	717	-4.0	42,600
1045	1439	407	-5.5	67,300	1180	1459	224	-5.2	124,900	1262	1468	717	-5.0	42,600
1047	1540	250	-4.2	109,200	1181	1431	223	-5.7	125,100	1263	1413	722	-6.0	42,400
1048	1576	635	-3.7	47,100	1182	1407	223	-6.1	125,200	1264	1340	717	-7.0	42,600
1049	1089	411	-10.4	66,700	1183	1383	224	-6.4	124,700	1265	1263	717	-8.0	42,600
1050	949	1040	-13.2	28,900	1184	1454	182	-5.3	164,400	1266	1182	720	-9.0	42,500
1051	426	818	-31.1	37,800	1185	1422	183	-5.8	162,600	1267	1110	717	-10.0	42,600
1052	1583	1385	-3.6	16,900	1186	1394	182	-6.3	164,300	1268	1055	717	-11.0	42,600
1053	779	1092	-16.8	27,000	1189	1171	214	-9.2	131,800	1269	999	717	-12.0	42,600
1054	1613	620	-3.2	48,000	1190	1457	286	-5.2	94,200	1270	959	715	-13.0	42,700
1055	1380	377	-6.5	72,000	1191	686	1114	-19.5	26,200	1271	905	712	-14.0	42,900
1056	284	663	<-35.0	45,500	1192	265	893	<-35.0	34,700	1272	857	714	-15.0	42,800
1058	1261	746	-8.0	41,200	1193	403	1292	-32.6	20,000	1273	810	705	-16.0	43,300
1060	393	605	-33.3	49,000	1194	344	1275	<-35.0	20,600	1274	774	711	-17.0	42,900
1061	1817	645	-0.9	46,600	1195	505	1311	-27.6	19,400	1277	737	708	-18.0	43,100
1062	1245	746	-8.2	41,200	1196	572	1293	-24.1	20,000	1278	702	711	-19.0	42,900
1064	1258	792	-8.1	39,000	1197	639	1502	-21.2	13,000	1279	671	710	-20.0	43,000
1065	705	934	-18.9	33,000	1198	637	1402	-21.3	16,300	1280	645	710	-21.0	43,000
1066	1181	734	-9.0	41,800	1199	614	1407	-22.1	16,200	1281	617	707	-22.0	43,100
1067	529	658	-26.3	45,800	1200	637	1431	-21.3	15,400	1282	595	704	-23.0	43,300
1068	508	696	-27.4	43,700	1201	1095	1394	-10.3	16,600	1283	573	700	-24.0	43,500
1069	1898	604	-0.3	49,100	1202	1719	1545	-2.1	11,600	1284	552	695	-25.0	43,700
1071	873	609	-14.7	48,700	1203	791	668	-16.5	45,200	1285	536	694	-26.0	43,800
1073	1768	1128	-1.5	25,800	1204	964	1021	-12.9	29,700	1286	515	687	-27.0	44,200
1075	836	773	-15.4	39,900	1205	313	195	<-35.0	148,700	1287	496	683	-28.0	44,400
1076	1863	861	-0.6	36,000	1208	306	194	<-35.0	149,800	1288	467	669	-29.0	45,200
1078	826	566	-15.7	51,600	1209	320	197	<-35.0	147,400	1289	447	667	-30.9	45,300
1081	971	483	-12.7	58,500	1210	326	197	<-35.0	146,600	1290	427	655	-31.0	45,900
1083	1697	202	-2.3	142,300	1211	394	294	-33.2	91,400	1291	412	655	-32.0	45,900
1085	1157	794	-9.4	38,900	1212	402	294	-32.7	91,200	1292	397	652	-33.0	46,100
1090	620	910	-21.9	34,000	1214	386	294	-33.7	91,400	1293	381	654	-34.0	46,000
1092	1867	597	-0.5	49,500	1215	641	329	-21.2	81,600	1294	365	653	-35.0	46,100
1093	2019	894	>0.0	34,600	1216	660	329	-20.4	81,600	1295	348	653	<-35.0	46,100
1094	1546	538	-4.1	53,700	1217	914	266	-13.8	101,800					
1095	1545	477	-4.1	59,100	1218	873	245	-14.7	112,000					
1098	61	935	<-35.0	33,000	1219	970	372	-12.7	72,900					
1099	1954	237	>0.0	116,000	1220	1021	298	-11.6	90,100					
1101	588	1048	-23.3	28,600	1221	1392	205	-6.3	139,500					
1102	1050	667	-11.1	45,200	1222	1354	203	-6.8	141,800					
1103	457	797	-29.5	38,800	1223	1362	205	-6.7	139,500					
1105	1884	532	-0.4	54,200	1224	673	540	-19.9	53,600					
1106	1714	649	-2.1	46,300	1225	614	542	-22.1	53,400					
1107	1717	546	-2.1	53,100	1226	603	539	-22.6	53,600					
1108	1976	722	>0.0	42,400	1227	696	623	-19.2	47,800					
1111	547	1066	-25.3	28,000	1228	707	628	-18.9	47,500					
1112	1348	621	-6.9	48,000	1229	475	447	-28.7	62,300					
1115	1385	762	-6.4	40,400	1230	466	1282	-29.0	20,400					
1116	1078	816	-10.6	38,000	1231	759	1461	-17.4	14,400					
1117	975	787	-12.6	39,300	1232	1324	1170	-7.2	24,200					
1118	1202	933	-8.7	33,100	1233	1583	1005	-3.6	30,300					
1119	1022	1076	-11.6	27,600	1234	1865	809	-0.6	38,200					
1120	1905	616	-0.3	48,300	1235	1812	817	-1.0	37,900					
1121	1512	1301	-4.5	19,700	1236	1411	703	-6.0	43,400					
1122	1114	677	-9.9	44,700	1237	1392	682	-6.3	44,500					
1123	1464	452	-5.1	61,700	1238	794	410	-16.4	66,900					
1125	1048	857	-11.1	36,200	1239	769	407	-17.1	67,300					
1126	1122	802	-9.8	38,600	1240	740	406	-17.9	67,500					
1128	1722	892	-2.1	34,700	1241	743	511	-17.8	55,900					
1133	1098	825	-10.2	37,500	1242	713	510	-18.7	56,000					
1139	1830	569	-0.8	51,400	1243	682	509	-19.6	56,100					
1147	764	1182	-17.3	23,800	1244	663	504	-20.3	56,500					
1148	1968	724	>0.0	42,300	1245	565	582	-24.4	50,500					

Table 2. Table of some identified proteins

POP name	Protein name	MSN's	Basis for identification
IDS:3_ALPHA_HDDH	3- $\alpha$ -hydroxysteroid-dihydrodiol-dehydrogenase, an enzyme of steroid metabolism	137, 159	Pure protein and antibody provided by Dr. T.M. Penning, Department of Pharmacology, School of Medicine, University of Pennsylvania.
IDS:ACTIN_BETA	$\beta$ cellular actin, a cytoskeletal protein	38	Homologous position with respect to other mammalian systems
IDS:ACTIN_GAMMA	$\gamma$ cellular actin, a cytoskeletal protein	68	Homologous position with respect to other mammalian systems
IDS:ALBUMIN	Serum albumin, mature form.	21, 28, 33	Predominance in rat plasma
IDS:APO_A-I	Apo A-I plasma lipoprotein, mature form (tentative).	236, 483	Presence in rat plasma, regulation by some lipid-lowering drugs
IDS:CALMODULIN	Calmodulin, an acidic cytosolic calcium-binding protein	123, 649	Homologous position with respect to other mammalian systems
IDS:CATALASE	Catalase (peroxisomal)	54, 61, 106	Presence in purified peroxisomes, similarity in position to mouse catalase
IDS:CPKSPOTS	Spots contributed by the CPK charge standards (not rat liver proteins)	1257 - 1295	
IDS:OPS	Carbamoyl phosphate synthase	114, 157, 167, 174, 1184, 1185, 1186, 1222	Pure protein provided by Dr. Margaret Marshall, Department of Pharmacology, Medical School, University of Wisconsin - Madison.
IDS:CYTOCHROME_B5	Cytochrome b5	87, 477	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center
IDS:FABP-L	Liver fatty-acid binding protein	227	Pure protein provided by Dr. Nathan Bass, Department of Medicine, University of California School of Medicine, San Francisco
IDS:HMG-COA_SYNTHASE	Cytosolic HMG-CoA Synthase	133, 144, 235, 413	Antibody provided by Dr. Michael Greenspan, Merck Sharp & Dohme Research Laboratories, Rahway, NJ
IDS:LAMIN_B	Lamin B, a nuclear protein	415, 734	Homologous position with respect to other mammalian systems
IDS:MITCON:1	Mitcon:1 (F1 ATPase $\beta$ subunit), a mitochondrial inner membrane	17, 49, 71, 340, 1245, 1246, 1247, 1249	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:MITCON:2	Mitcon:2, a mitochondrial matrix stress protein equivalent to E.	15, 25, 110, 1241, 1242, 1243, 1244	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:MITCON:3	Mitcon:3, a mitochondrial matrix stress protein, likely analog of	18, 35, 226, 600, 1238, 1239, 1240	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:NADPH_P450_RED	NADPH cytochrome P-450 reductase, frequently co-induced with P-450's	175, 251, 812	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center
IDS:PDI	Protein disulphide isomerase 1	168, 1170, 1171, 1172	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis
IDS:PLASMA_PROTEINS	Rat plasma proteins observed in liver	21, 28, 33, 44, 72, 102, 115, 197, 236, 246, 248, 257, 293, 332, 347, 364, 369, 419, 432, 463, 468, 518, 562, 605, 623, 666, 667, 725, 738, 790, 865, 903, 926	Plasma coelectrophoresis studies
IDS:PRO-ALBUMIN	Serum albumin precursor	47, 93	Relative position to mature albumin, presence in micro-somes
IDS:PYR-CARBOX	Pyruvate carboxylase	179, 1180, 1181, 1182, 1183	Pavlica, R.J., et al. BBA (1990) 1022:115-125.
IDS:SOD	Superoxide dismutase	135	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis
IDS:TUBULIN_ALPHA	$\alpha$ tubulin, a cytoskeletal protein	56, 132, 1224, 1252	Homologous position with respect to other mammalian systems
IDS:TUBULIN_BETA	$\beta$ tubulin, a cytoskeletal protein	50, 1225, 1226, 1251	Homologous position with respect to other mammalian systems

e 3. Computed pI's of two sets of carbamylated protein standards: Rabbit muscle CPK and human hemoglobin (Hb)

Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	NH2- 7.0	Calc pI	Real CPK
Rabbit muscle CPK	KIRBCM	28	27	17	34	18	1	6.84	0.0
		28	27	17	33	18	1	6.67	-1
		28	27	17	32	18	1	6.54	-2
		28	27	17	31	18	1	6.42	-3
		28	27	17	30	18	1	6.31	-4
		28	27	17	29	18	1	6.21	-5
		28	27	17	28	18	1	6.12	-6
		28	27	17	27	18	1	6.03	-7
		28	27	17	26	18	1	5.94	-8
		28	27	17	25	18	1	5.85	-9
		28	27	17	24	18	1	5.76	-10
		28	27	17	23	18	1	5.67	-11
		28	27	17	22	18	1	5.58	-12
		28	27	17	21	18	1	5.48	-13
		28	27	17	20	18	1	5.39	-14
		28	27	17	19	18	1	5.29	-15
		28	27	17	18	18	1	5.20	-16
		28	27	17	17	18	1	5.12	-17
		28	27	17	16	18	1	5.04	-18
		28	27	17	15	18	1	4.96	-19
		28	27	17	14	18	1	4.89	-20
		28	27	17	13	18	1	4.83	-21
		28	27	17	12	18	1	4.77	-22
		28	27	17	11	18	1	4.71	-23
		28	27	17	10	18	1	4.66	-24
		28	27	17	9	18	1	4.61	-25
		28	27	17	8	18	1	4.56	-26
		28	27	17	7	18	1	4.52	-27
		28	27	17	6	18	1	4.48	-28
		28	27	17	5	18	1	4.44	-29
		28	27	17	4	18	1	4.40	-30
		28	27	17	3	18	1	4.36	-31
		28	27	17	2	18	1	4.32	-32
		28	27	17	1	18	1	4.29	-33
		28	27	17	0	18	1	4.25	-34
		28	27	17	0	18	0	4.22	-35
Hb-beta, human	HBHU	7	8	9	11	3	1	7.18	
		7	8	9	10	3	1	6.79	
		7	8	9	9	3	1	6.53	-1.8
		7	8	9	8	3	1	6.32	-3.2
		7	8	9	7	3	1	6.13	-5.3
		7	8	9	6	3	1	5.96	-7.2
		7	8	9	5	3	1	5.78	-10.0
		7	8	9	4	3	1	5.59	-12.3
		7	8	9	3	3	1	5.37	-15.5
		7	8	9	2	3	1	5.14	-18.0
		7	8	9	1	3	1	4.91	-21.0
		7	8	9	0	3	1	4.71	-25.5
		7	8	9	0	3	0	4.54	-27.2



Table 4. Computed pI's of some known proteins related to measured CPK pI's

Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	Calc pI	Real CPK
0 Creatine phospho kinase (CPK), rabbit muscle	KIRBCM	28	27	17	34	18	6.84	0.0
1 Fatty acid-binding protein, rat hepatic	FZRTL	5	13	2	16	2	7.83	-3.0
2 b2-microglobulin, human	MGHUB2	7	8	4	8	5	6.09	-5.0
3 Carbamoyl-phosphate synthase, rat	SYRTCA	72	96	28	95	56	5.97	-5.5
4 Proalbumin ( serum albumin precursor), rat	ABRTS	32	57	15	53	27	5.98	-6.2
5 Serum albumin, rat	ABRTS	32	57	15	53	24	5.71	-9.0
6 Superoxid dismutase (Cu-Zn, SOD), rat	A26810	8	11	10	9	4	5.91	-9.2
7 Phospholipase C, phosphoinositide-specific (?), rat	A28807	34	42	9	49	21	5.92	-9.2
8 Albumin, human	ABHUS	36	61	16	60	24	5.70	-11.9
9 Apo A-I lipoprotein, rat	A24700	18	24	6	23	12	5.32	-13.7
10 proApo A-I lipoprotein, human	LPHUA1	16	30	6	21	17	5.35	-14.3
11 NADPH cytochrome P-450 reductase, rat	RDRTO4	41	60	21	38	36	5.07	-15.6
12 Retinol binding protein, human	VAHU	18	10	2	10	14	5.04	-16.9
13 Actin beta, rat	ATRTC	23	26	9	19	18	5.06	-17.2
14 Actin gamma, rat	ATRTC	20	29	9	19	18	5.07	-16.6
15 Apo A-I lipoprotein, human	LPHUA1	16	30	5	21	16	5.10	-17.5
16 Apo A-IV lipoprotein, human	LPHUA4	20	49	8	28	24	4.88	-19.7
17 Tubulin alpha, rat	UBRTA	27	37	13	19	21	4.66	-19.8
18 F1ATPase beta, bovine	PWBOB	25	36	9	22	22	4.80	-21.0
19 Tubulin beta, pig	UBPGB	26	36	10	15	22	4.49	-22.5
20 Protein disulphide isomerase (PDI), rat hepatic	ISRTSS	43	51	11	51	9	4.07	-25.0
21 Cytochrome b5, rat	CBRT5	10	15	6	10	4	4.59	-26.0
22 Apo C-II lipoprotein, human	LPHUC2	4	7	0	6	1	4.44	-30.5
Amino acid pI assumed in calculation:		3.9	4.1	6.0	10.8	12.5		

Wirth  
Luo  
Fujimoto  
C. Bisgaard  
D. Olson  
Laboratory of Experimental  
Hepatic Cancer  
National Cancer Institute  
Bethesda, MD

# Contents

Introduction ..  
Materials and  
Methods ..  
1. Cells ..  
2. Metabolic  
3. Nine and  
4. Sample p  
5. Subcellul  
6. 2-D PAG  
7. Computer  
8. retograms  
9. Results ..  
10. [<sup>35</sup>S]Methi  
11. Whole ce  
12. Subcellul  
13. [<sup>32</sup>P]Ortho  
14. Discussion ..  
15. References ..  
16. Addendum 1:  
17. Addendum 2:  
18. Proteins ..

Correspondence: Dr. P  
National Car  
USA

Abbreviations: 2-D PA  
HLE, hum  
weight; NE  
10, Nonidet P-  
side; RLE, rat

Verlagsgesellschaft

N. Leigh Anderson<sup>1</sup>  
 Ricardo Esquer-Blasco<sup>1</sup>  
 Jean-Paul Hofmann<sup>1</sup>  
 Lydie Meheus<sup>2</sup>  
 Jos Raymackers<sup>2</sup>  
 Sandra Steiner<sup>3</sup>  
 Frank Witzmann<sup>4</sup>  
 Norman G. Anderson<sup>1</sup>

<sup>1</sup>Large Scale Biology Corporation,  
 Rockville, MD

<sup>2</sup>Innogenetics NV, Ghent

<sup>3</sup>Sandoz Pharma Ltd, Drug Safety  
 Assessment, Toxicology, Basel

<sup>4</sup>Molecular Anatomy Laboratory,  
 Indiana University Purdue  
 University Columbus, Columbus, IN

## An updated two-dimensional gel database of rat liver proteins useful in gene regulation and drug effect studies

We have improved upon the reference two-dimensional (2-D) electrophoretic map of rat liver proteins originally published in 1991 (N. L. Anderson *et al.*, *Electrophoresis* 1991, 12, 907-930). A total of 53 proteins (102 spots) are now identified, many by microsequencing. In most cases, spots cut from wet, Coomassie Blue stained 2-D gels were submitted to internal tryptic digestion [2], and individual peptides, separated by high-performance liquid chromatography (HPLC), were sequenced using a Perkin-Elmer 477A sequencer. Additional spots were identified using specific antibodies.

Figure 1 shows the current annotated 2-D map of F344 rat liver, analyzed using the Iso-DALT system (20 × 25 cm gels) and BDH 4-8 carrier ampholytes. Both the map itself and the master spot number system remain the same as shown in the original publication. Table 1 lists the important features of each identification shown, including the gel position, *pI*, and *M<sub>r</sub>*, for the most abundant or most basic form of each protein. Using this extended base of identified spots, a series of four improved calibration functions has been derived for the *pI* and SDS-*M<sub>r</sub>* axes (the first two of which are shown in Fig. 2A and B). Both forward and reverse functions are derived, so that one can compute the physical properties of a spot with a given gel location, or inversely compute the gel position expected for a protein having given physical properties:

$$Y_{\text{RATLIVER}} = f_{M-\text{RATLIVER}}(M_{\text{SEQUENCE-DERIVED}}) \quad (1)$$

$$X_{\text{RATLIVER}} = f_{pI-\text{RATLIVER}}(pI_{\text{SEQUENCE-DERIVED}}) \quad (2)$$

$$M_{\text{GEL-DERIVED}} = f_{\text{RATLIVER } Y-M_r}(Y_{\text{RATLIVER}}) \quad (3)$$

$$pI_{\text{GEL-DERIVED}} = f_{\text{RATLIVER } X-pI}(X_{\text{RATLIVER}}) \quad (4)$$

A spreadsheet program (in Microsoft Excel) was developed to facilitate flexible computation of *pI*'s from amino acid sequence data, and the results were entered into a relational database (Microsoft Access). A table of spot positions and sequence-derived *pI*'s and *M<sub>r</sub>*'s was fitted with a large series of analytic equations using Tablecurve (Jandel Scientific), and the four conversion Eqs. (1)-(4), relating computed *pI* and gel *X* coordinate, or computed molecular weight and gel *Y* coordinate, were selected, based on criteria of simplicity, goodness of fit and favorable asymptotic behavior. Table 2 lists the equations and coefficients. Application of Eqs. (3) and (4) to a spot's *X* and *Y* coordinates, given in [1], produce improved *M<sub>r</sub>* estimates, and allow computation of *pI*

directly in *pH* units, instead of in terms of positions relative to creatine phosphokinase (CPK) charge standards. The inverse Eqs. (1) and (2) were used to compute the gel positions of a series of *pI* and *M<sub>r</sub>* tick marks. These tick marks were plotted with SigmaPlot (Jandel), together with fiducial marks locating several prominent spots, and the resulting graphic was aligned over the synthetic gel image (computed by Kepler from the master gel pattern) using Freelance (Lotus Development). Maps were printed as Postscript output from Freelance, either in black and white (as shown here) or in color, where label color indicates subcellular location (available from the first author upon request). We have also used the rat liver 2-D pattern as presented here to calibrate the patterns of other samples. Using mixtures of rat liver and mouse liver samples, for example, we made composite 2-D patterns that allow use of the rat pattern to standardize both axes of the mouse pattern. This was accomplished by deriving transformations relating the rat and mouse *X*, and separately the rat and mouse *Y*, axes (Table 2, lower half; Fig. 2C and D) based on a series of spots that coelectrophorese in these closely related species. These functions were then applied to derive equations relating the mouse liver *X* and *Y* to *pI* and SDS-*M<sub>r</sub>* (Eqs. 5 and 6 below). The resulting standardized 2-D pattern for B6C3F1 mouse liver is shown in Fig. 3.

$$M_{\text{MOUSELIVER}} = f_{\text{RATLIVER } Y-M_r}(f_{\text{MOUSELIVER } Y-\text{RATLIVER } Y}(Y_{\text{MOUSELIVER}})) \quad (5)$$

$$pI_{\text{MOUSELIVER}} = f_{\text{RATLIVER } X-pI}(f_{\text{MOUSELIVER } X-\text{RATLIVER } X}(X_{\text{MOUSELIVER}})) \quad (6)$$

A slightly more complex approach can be used to standardize samples that have few or no spots co-electrophoresing with rat liver proteins. In this case, a 2-D gel is prepared with a mixture of the two samples, and four functions (forward and backward, each for *X* and *Y*) are derived relating each sample's own master pattern to the composite. The required functions are then applied in a nested fashion to yield the desired result (using rat plasma as an example):

$$M_{\text{RATPLASMA}} = f_{\text{RATLIVER } Y-M_r}(f_{\text{RATPLASMA+LIVER } Y-\text{RATLIVER } Y}(f_{\text{RATPLASMA } Y-\text{RATPLASMA+LIVER } Y}(Y_{\text{RATPLASMA}})))) \quad (7)$$

Correspondence: Dr. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338 USA (Tel: +301-424-5989; Fax: +301-762-4892; email: leigh@lsbc.com)

Keywords: Two-dimensional polyacrylamide gel electrophoresis / Liver / Map / Identification / Calibration

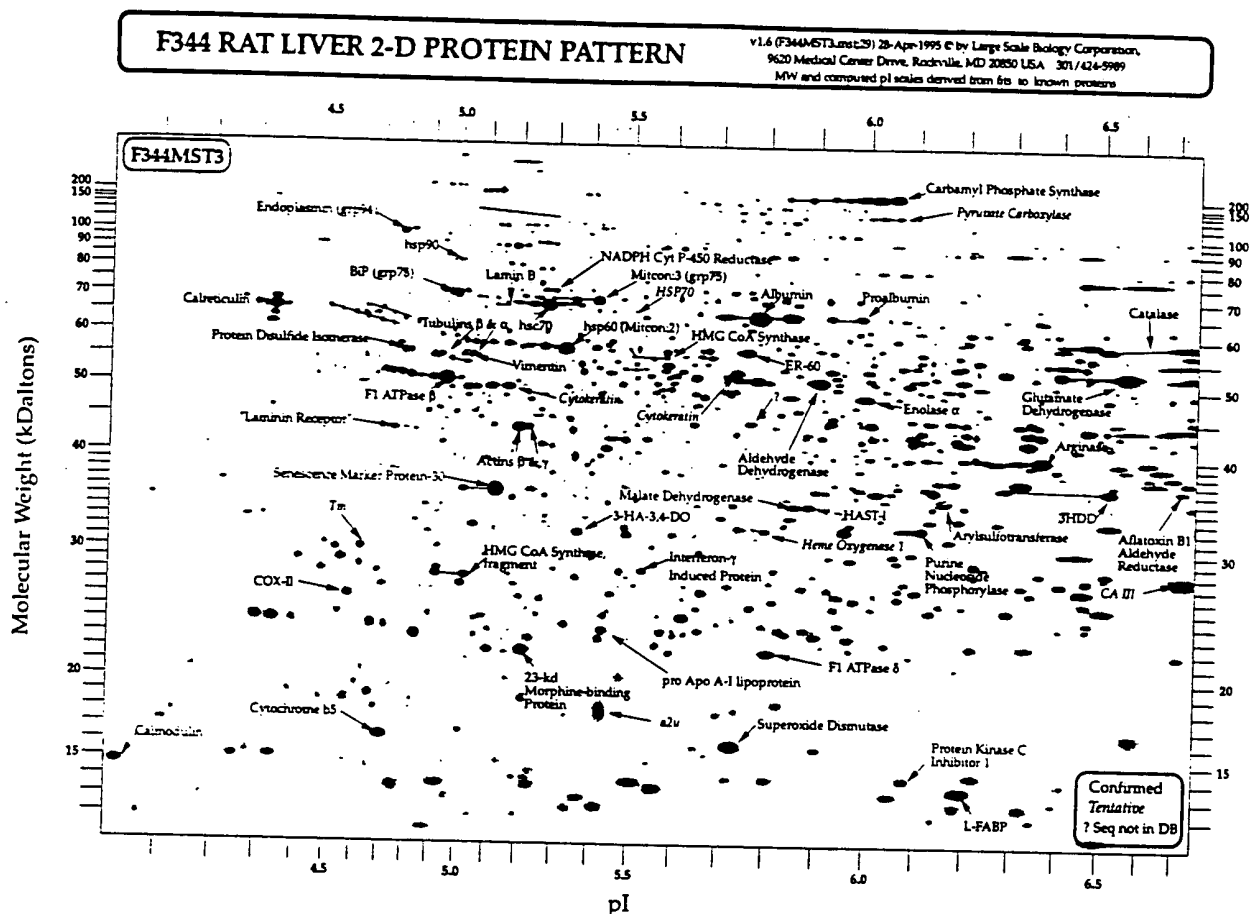


Figure 1. Master 2-D gel pattern of Fischer 344 rat liver proteins, annotated with 53 protein identifications and computed  $pI$  and  $M_r$  axes. Tentative identifications are in italic type.

Table 1. Proteins identified in the 2-D pattern of F344 rat liver

MSN <sup>a)</sup>	Protein ID <sup>b)</sup>	Protein name	Identification comments	Gel X <sup>c)</sup>	Experimental $pI$ <sup>d)</sup>	Gel Y <sup>c)</sup>	Experimental $M_r$ <sup>d)</sup>
126	HADO-HUMAN <sup>e)</sup>	3-HA-3,4-DO: 3-hydroxy-anthranilate-3,4-dioxygenase	Internal sequence	871.95	5.36	921.35	30 207
137, 159, 288, 258	DIDH_RAT	3HDD: 3-hydroxysteroid dihydrodiol reductase	Ab (T.M. Penning) and pure protein	1857.52	6.51	822.52	34 406
173	MUP_RAT	$\alpha_2u$ globulin	Presence in liver microsome lumen, abundance in kidney, $pI$ , $M_r$	919.16	5.43	1313.81	19 549
38	ACTB_HUMAN	Actin $\beta$	Analogy with other mammalian patterns (e.g. human) through coelectrophoresis	763.40	5.19	693.64	41 586
68	ACTG_HUMAN	Actin $\gamma$	Analogy with other mammalian patterns (e.g. human) through coelectrophoresis	779.42	5.21	692.26	41 677
693	AFAR_RAT	Aflatoxin B1 aldehyde reductase	Internal sequence	1993.32	6.72	818.60	34 593
28, 21, 33	ALBU_RAT	Albumin	Coelectrophoresis with principal plasma protein	1262.81	5.86	445.64	66 354
43	DHAM_RAT	Aldehyde dehydrogenase	N-Terminal sequence and AAA	1317.72	5.91	589.03	49 602
96	ARGI_RAT	Arginase	Internal sequence	1730.72	6.34	756.02	37 819
117	SUAR_RAT	Arylsulfotransferase	Internal sequence	1547.96	6.14	849.08	33 186
1163, 1161, 1162, 20	GR78_RAT	BIP (GRP-78)	Ab (F. Witzmann)	665.33	5.01	397.39	74 564
185	CAH3_RAT	CA-III	Uncertain; by comparison with mouse	1996.60	6.72	1017.02	26 887
123	CALM_HUMAN	Calmodulin	Analogy with human cellular patterns through coelectrophoresis	23.05	4.03	1433.25	17 419
3, 201, 48, 39, 22, 24	CRTC_RAT	Calreticulin	Ab (Lance Pohl)	310.59	4.34	433.80	68 206

Table 1. continued

MSN <sup>a)</sup>	Protein IDb)	Protein name	Identification comments	Gel X <sup>c)</sup>	Experimental pI <sup>d)</sup>	Gel Y <sup>c)</sup>	Experimental M <sub>r</sub> <sup>d)</sup>
1184, 1186, 114, 174, 118, 5, 167, 157, 54, 61, 136	CPSM_RAT CATA_RAT COX2_RAT	Carbamyl phosphate synthase Catalase COX-II	2-D of pure protein; confirmed by N-terminal sequence and AAA Internal sequence Ab (J. W. Taanman), confirmed by internal sequence	1453.56 2000.81 452.57	6.05 6.73 4.61	181.64 499.64 1062.67	160 640 58 968 25 504
87	CYB5_RAT	Cytochrome B5	2-D of pure protein; Ab; confirmed by AAA	515.68	4.73	1370.55	18 493
41	CK-RAT <sup>e)</sup>	Cytokeratin	Location in cytoskeletal fraction	1165.12	5.75	569.09	51 448
29	CK-RAT <sup>e)</sup>	Cytokeratin	Location in cytoskeletal fraction	743.11	5.15	605.23	48 187
5, 11	ENPL-RAT <sup>e)</sup>	Endoplasmic	Ab (F. Witzmann)	567.73	4.83	263.37	112 194
60	ENOA_RAT	Enolase A	Internal sequence and AAA	1399.78	6.00	623.54	46 674
27	ER60_RAT	ER-60	N-Terminal sequence (R. M. Van Frank)	1184.20	5.77	523.51	56 169
17	ATPB_RAT	F1 ATPase $\beta$	N-Terminal sequence and AAA	629.06	4.95	588.83	49 620
196	ATP7_RAT	F1 ATPase $\delta$	Internal sequence	1227.24	5.82	1184.65	22 310
79	F16P_RAT	Fructose-1,6-bis-phosphatase	Uncertain; by comparison with ID in Garrison and Wager (JBC 257:13135-13143)	924.54	5.44	737.77	38 858
62, 78	DHE3_RAT	Glutamate dehydrogenase	N-Terminal sequence and internal sequence	1887.39	6.55	566.92	51 655
125	HAST-RAT <sup>e)</sup>	HAST-I: N-hydroxyaryl-amine sulfotransferase	Internal sequence	1297.94	5.89	861.55	32 638
307	HO1_RAT	Heme oxygenase 1	Uncertain; available data from internal sequence	1219.39	5.81	915.71	30 423
413, 1250, 933	HMCS_RAT	HMG CoA synthase, cytosolic	Ab (J. Germershausen)	1033.48	5.59	538.13	54 571
133, 144, 235	HMCS_RAT	HMG CoA synthase, mitochondrial (frag)	Ab (J. Germershausen), N-terminal sequence (Steiner/Lottspeich)	666.40	5.02	1019.42	26 811
8, 23, 1307	HS7C_RAT	HSC-70	Positional homology (with human, etc.) through coelectrophoresis	811.87	5.27	425.76	69 521
15, 25, 110	P60_RAT	HSP-60	Ab (F. Witzman); confirmed by N-terminal sequence and AAA	845.09	5.32	520.03	56 561
971	HS70-RAT <sup>e)</sup>	HSP-70	Ab (F. Witzman)	976.11	5.51	437.14	67 674
1216, 1215, 90	HS90-RAT <sup>e)</sup>	HSP-90	Ab (F. Witzman)	659.86	5.00	329	90 107
256	INGI-HUMAN	Interferon- $\gamma$ induced protein	Internal sequence	993.85	5.54	1006.04	27 237
415, 734	LAMB-RAT <sup>e)</sup>	Lamin B	Positional homology with human through coelectrophoresis, nuclear location	737.10	5.14	425.19	69 615
80	LAMR-RAT <sup>e)</sup>	"Laminin receptor"	Internal sequence	534.02	4.77	697.62	41 327
227	FABL_RAT	L-FABP (liver fatty acid binding protein)	Ab (N. M. Bass)	1586.09	6.18	1483.43	16 622
134	MDHC_MOUSE E	Malate dehydrogenase	Internal sequence	1270.85	5.86	861.96	32 620
18, 35, 226	GR75-RAT <sup>e)</sup>	Mitcon-3; grp75	Positional homology with human through coelectrophoresis	905.67	5.41	413.67	71 589
175, 251	NCPR_RAT	NADPH P450 reductase	2-D of pure protein	824.69	5.29	393.21	75 366
1168, 1170, 1171	PDI_RAT	PDI: Protein disulfide isomerase	N-Terminal sequence (R. M. van Frank), Ab	564.30	4.83	528.47	55 618
47, 93	ALBU_RAT	Pro-Albumin	Microsomal lumen location, pI, M <sub>r</sub> relative to albumin	1391.03	5.99	446.68	66 195
236	APA1_RAT	Pro-APO A-I lipoprotein	Coelectrophoresis with plasma protein	920.41	5.43	1137.51	23 467
320	IPK1_BOVIN	Protein kinase C inhibitor 1	Internal sequence; homology with bovine protein	1480.01	6.08	1458.81	17 007
152	PNPH_MOUSE	Purine nucleoside phosphorylase	Internal sequence	1507.19	6.10	911.16	30 599
1179, 1180, 1181, 1182, 1183	PYVC-RAT <sup>e)</sup>	Pyruvate carboxylase	Tentative; 2-D of pure protein (J. G. Henslee, JBC, 1979); reported in <i>Biochim. Biophys. Acta</i> 1022, 115-125	1485.10	6.08	223.52	131 589
55, 103	SM30_RAT	SMP-30: Senescence marker protein-30	Internal sequence	721.71	5.11	830.10	34 051
135	SODC_RAT	Superoxide dismutase	AAA; confirmed by internal sequence (R. M. Van Frank)	1161.24	5.74	1388.68	18 173
172	TPM-RAT <sup>e)</sup>	Tm: tropomyosin	Location in cytoskeleton, 2-D position relative to human, Ab	476.24	4.66	957.86	28 865
277, 56	TBA1_RAT	Tubulin $\alpha$	Positional homology with human through coelectrophoresis, cytoskeletal location	688.22	5.06	537.67	54 620
50, 1225	TBB1_RAT	Tubulin $\beta$	Positional homology with human through coelectrophoresis, cytoskeletal location	621.29	4.93	535.48	54 855
1224	VIME_RAT	Vimentin	Positional homology with human through coelectrophoresis, cytoskeletal location	673.00	5.03	539.50	54 426

Table 1. continued

MSN <sup>a)</sup>	Protein IDb)	Protein name	Identification comments	Gel X <sup>c)</sup>	Experimental pI <sup>d)</sup>	Gel Y <sup>c)</sup>	Experimental M <sub>r</sub> <sup>d)</sup>
113	Unknown	? not in sequence databases	Internal sequence	1191.28	5.78	680.42	42 469
104	BBPL_RAT	23 kDa morphine-binding protein	Internal sequence	773.31	5.20	1182.41	22 363

a) Master spot number (MSN) from [1]

b) SwissPROT identifier

c) Coordinates of the most basic or most abundant assigned spot on the F344 master gel pattern

d) pI and M<sub>r</sub> of the most basic or most abundant assigned spot, derived from the calibration functions included here

e) SwissPROT style proposed identifier

Abbreviations: AAA, amino acid analysis; Ab, antibody

Table 2. Equations and coefficients

Function	Equation (f)	r <sup>2</sup>	a	b	c	d	e
Rat gel Y = f(computed M <sub>r</sub> )	$y = a + b \exp(-x/c)$	0.988181021	178.74803	1967.7892	32363.958		
Rat gel X = f(computed pI)	$y = a + bx + cx/\ln x + d/x + e/x^{1.5}$	0.99247216	-8685665.5	-904497.94	3856926.1	18276844	-27154534
Computed M <sub>r</sub> = f(rat gel Y)	$y = a + bxc$	0.9960177	-8464.5809	19095881	-0.9086255		
Computed pI = f(rat gel X)	$y = a + bx + cx^2 + dx^2 \ln x + ex^3$	0.99176499	4.044686	-0.00114238	0.0000323	-0.00000455	0.0000000176
Mouse gel Y = f(rat gel Y)	$y = a + bx + cx^{1.5} + dx^{0.5} \ln x + ex/\ln x$	0.99951069	11861.44	678.91666	-0.78964914	1567.5639	-6953.9592
Mouse gel X = f(rat gel X)	$y = a + bx^2 \ln x + cx^{2.5} + dx^3$	0.99926349	58.935923	0.00091353	-0.000213688	0.00000159	
Rat gel Y = f(mouse gel Y)	$y = a + bx^2 \ln x + cx^{2.5} + dx^3$	0.99950032	69.740526	0.00050772	-0.000130392	0.00000116	
Rat gel X = f(mouse gel X)	$y = a + bx + cx^2 \ln x + dx^{2.5} + ex^3$	0.9992832	-198.07189	2.0899063	-0.000671191	0.000145189	-0.000000986

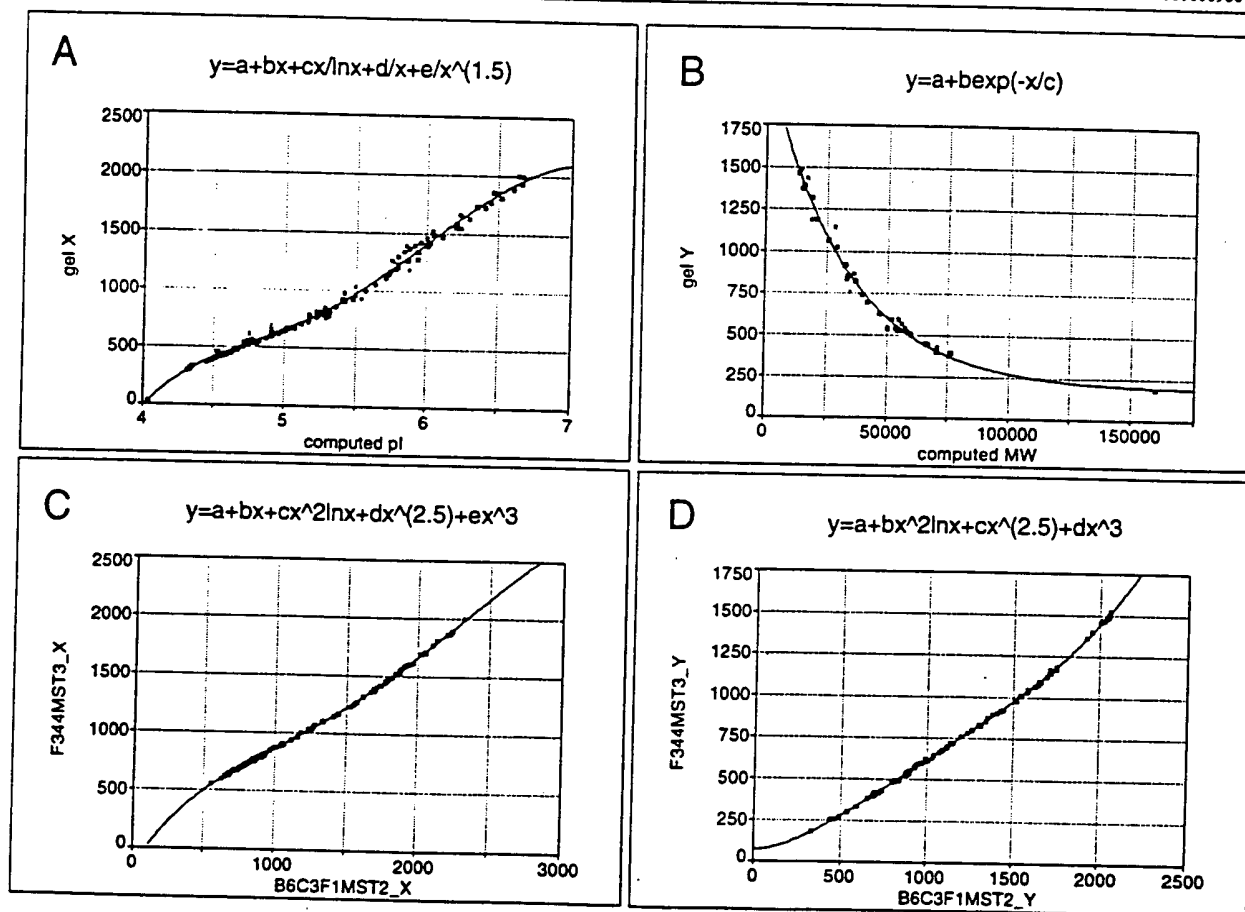


Figure 2. Plots showing fits of selected equations (continuous curves) to data on identified proteins (square symbols). (A) pI/computed from sequence data versus gel X position for identified spots in F344 rat liver; (B) M<sub>r</sub>, computed from sequence data versus gel Y position for identified spots in F344 rat liver; (C) gel X position for spots in B6C3F1 mouse liver versus X position in F344 rat liver, for coelectrophoresing spots; (D) gel Y position for spots in B6C3F1 mouse liver versus Y position in F344 rat liver, for coelectrophoresing spots. In each case, inverse equations were also computed (Table 2).

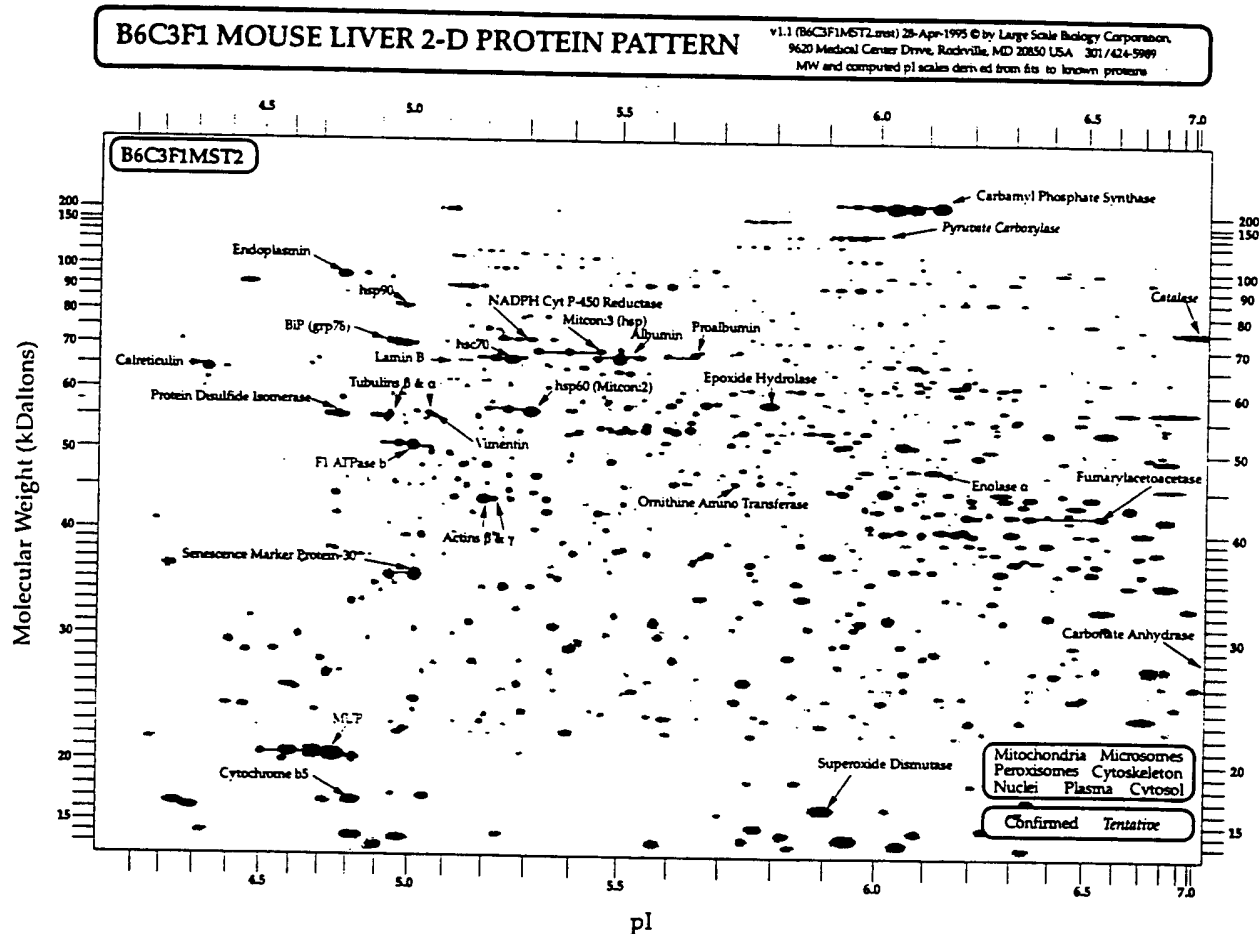


Figure 3. Master 2-D gel pattern for B6C3F1 mouse liver, standardized using the F344 rat liver pattern identifications, according to the method described in the text. Twenty-nine proteins are identified.

$$pI_{\text{RATPLASMA}} = \frac{\int_{\text{RATLIVER}} X - pI (\int_{\text{RATPLASMA} + \text{LIVER}} X - \text{RATLIVER} X)}{\int_{\text{RATPLASMA} + \text{LIVER}} X - \text{RATPLASMA} + \text{LIVER} X (X_{\text{RATPLASMA}}))} \quad (8)$$

This unified approach, in which one well-populated 2-D pattern is used to standardize a family of other patterns, has the additional advantage that the resulting  $pI$  and  $M_r$  scales are directly compatible. Hence one can compare the relative  $pI$ 's of mouse and rat versions of a sequenced protein in a consistent  $pI$  measurement system, and select likely inter-species analogs based on positional relationships on common scales. Adoption of immobilized pH gradient (IPG) technology [4-7] will result in substantial improvements in  $pI$  positional reproducibility for standard 2-D maps such as those presented here; however, we believe that our approach will continue to be useful in establishing the empirical pH gradient actually achieved by such gels under given experimental conditions (temperature, urea concentration, etc.), in relating patterns run on different IPG ranges and using different lots of IPG gels (between which some variation will persist). Development of rodent organ maps is a continuing effort in our laboratories [8-10], and results in regular additions of identified proteins. Those who wish to receive current rodent liver maps, with color annotations, should send a stamped self-addressed envelope to the first author.

We would like to thank the individuals who provided antibodies mentioned in Table 1, and R. M. van Frank for unpublished sequenced data.

Received May 31, 1995

## References

- [1] Anderson, N. L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, 12, 907-930.
- [2] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, 203, 173-179.
- [3] Witzmann, F., Clack, J., Fultz, C., Jarnot, B., *Electrophoresis* 1995, 16, 451-459.
- [4] Rosengren, A. E., Bjellqvist, B., Gasparic, V., *US Patent* 4130470, December 1978.
- [5] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, 4, 321-326.
- [6] Görg, A., Postel, W., Günther, S., Weser, J., *Electrophoresis* 1985, 6, 599-604.
- [7] Gianazza, E., Astrua-Testori, S., Giacon, P., Righetti, P. G., *Electrophoresis* 1985, 6, 332-339.
- [8] Myers, T. G., Dietz, E. C., Anderson, N. L., Khairallah, E. A., Cohen, S. D., Nelson, S. D., *Chem. Res. Toxicol.* 1995, 8, 403-413.
- [9] Cunningham, M. L., Pippin, L. L., Anderson, N. L., Wenk, M. L., *Toxicol. Appl. Pharmacol.* 1995, 131, 216-223.
- [10] Anderson, N. L., Copple, D. C., Bende, R. A., Probst, G. S., Richardson, F. C., *Fundam. Appl. Toxicol.* 1992, 18, 570-580.

# Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It

MARC R. WILKINS<sup>1</sup>, JEAN-CHARLES SANCHEZ<sup>1</sup>, ANDREW. A. GOOLEY<sup>1</sup>,  
RON D. APPEL<sup>3</sup>, IAN HUMPHERY-SMITH<sup>2</sup>, DENIS F. HOCHSTRASSER<sup>3</sup>  
AND KEITH L. WILLIAMS<sup>1,\*</sup>

<sup>1</sup> Macquarie University Centre for Analytical Biotechnology, Macquarie University, Sydney, NSW 2109, Australia; <sup>2</sup> Department of Microbiology, University of Sydney, NSW, 2006, Australia and <sup>3</sup> Central Clinical Chemistry Laboratory and Medical Computing Centre of the University of Geneva, CH 1211 Geneva 14, Switzerland

## Introduction

The advent of large genome sequencing projects has changed the scale of biology. Over a relatively short period of time, we have witnessed the elucidation of the complete nucleotide sequence for bacteriophage  $\lambda$  (Sanger *et al.*, 1982), the nucleotide sequence of an eukaryotic chromosome (Oliver *et al.*, 1992), and in the near future will see the definition of all open reading frames of some simple organisms, including *Mycoplasma pneumoniae*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. Nevertheless, genome sequencing projects are not an end in themselves. In fact, they only represent a starting point to understanding the function of an organism. A great challenge that biologists now face is how the co-expression of thousands of genes can best be examined under physiological and pathophysiological conditions, and how these patterns of expression define an organism.

There are two approaches that can be used to examine gene expression on a large scale. One uses nucleic acid-based technology, the other protein-based technology. The most promising nucleic-acid based technology is differential display of mRNA (Liang and Pardee, 1992; Bauer *et al.*, 1993), which uses polymerase chain reaction with arbitrary primers to generate thousands of cDNA species, each which correspond to an expressed gene or part of a gene. However, it is currently unclear if this technique can be developed to reliably assay the expression of thousands of genes or

---

\* Corresponding Author

identify all cDNA species, and the approach does not easily allow a systematic screening. Analysis of gene expression by the study of proteins present in a cell or tissue presents a favorable alternative. This can be achieved by use of two-dimensional (2-D) gel electrophoresis, quantitative computer image analysis, and protein identification techniques to create 'reference maps' of all detectable proteins. Such reference maps establish patterns of normal and abnormal gene expression in the organism, and allow the examination of some post-translational protein modifications which are functionally important for many proteins. It is possible to screen proteins systematically from reference maps to establish their identities.

To define protein-based gene expression analysis, the concept of the 'proteome' was recently proposed (Wilkins *et al.*, 1995; Wasinger *et al.*, 1995). A proteome is the entire PROTEin complement expressed by a genOME, or by a cell or tissue type. The concept of the proteome has some differences from that of the genome, as while there is only one definitive genome of an organism, the proteome is an entity which can change under different conditions, and can be dissimilar in different tissues of a single organism. A proteome nevertheless remains a direct product of a genome. Interestingly, the number of proteins in a proteome can exceed the number of genes present, as protein products expressed by alternative gene splicing or with different post-translational modifications are observed as separate molecules on a 2-D gel. As an extrapolation of the concept of the 'genome project', a 'proteome project' is research which seeks to identify and characterise the proteins present in a cell or tissue and define their patterns of expression.

Proteome projects present challenges of a similar magnitude to that of genome projects. Technically, the 2-D gel electrophoresis must be reproducible and of high resolution, allowing the separation and detection of the thousands of proteins in a cell. Low copy number proteins should be detectable. There should be computer gel image analysis systems that can qualitatively and quantitatively catalog the electrophoretically separated proteins, to form reference maps. A range of rapid and reliable techniques must be available for the identification and characterisation of proteins. As a consequence of a proteome project, protein databases must be assembled that contain reference information about proteins: such databases must be linked to genomic databases and protein reference maps. Databases should be widely accessible and easy to use.

Recently, there have been many changes in the techniques and resources available for the analysis of proteomes. It is the aim of this chapter to discuss the status of the areas outlined above, and to review briefly the progress of some current proteome projects.

### Two-dimensional electrophoresis of proteomes

Two dimensional (2-D) gel electrophoresis involves the separation of proteins by their isoelectric point in the first dimension, then separation according to molecular weight by sodium dodecyl sulfate electrophoresis in the second dimension. Since first described (Klose, 1975; O'Farrell, 1975; Scheele, 1975), it has become the method of choice for the separation of complex mixtures of proteins, albeit with many modifications to the original techniques. 2-D electrophoresis forms the basis of proteome projects through separating proteins by their size and charge (Hochstrasser *et al.*,

Fig:  
illu-  
was  
(top  
the  
of k

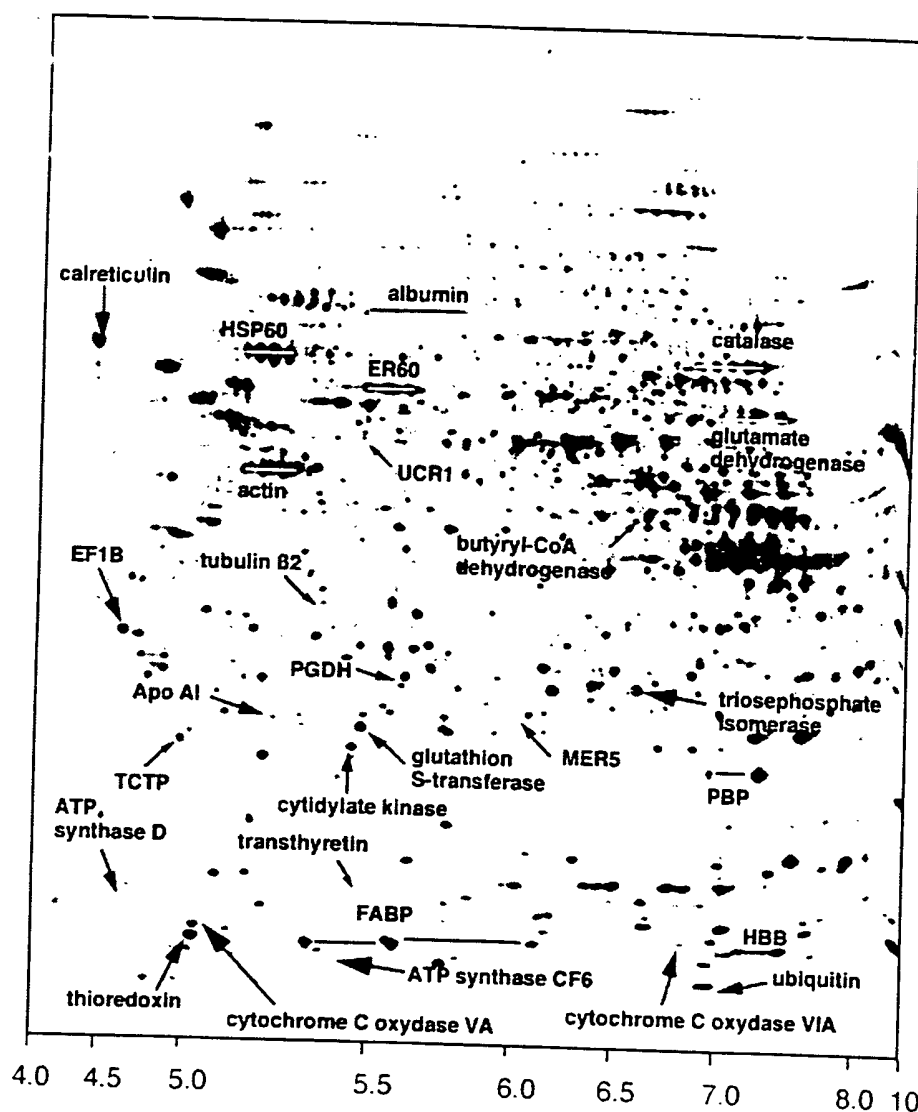
19  
pro  
sin

2-I

A  
ph  
en



## HEPG2 2D-PAGE MAP



**Figure 1.** Two-dimensional gel electrophoresis map of a human hepatoblastoma-derived cell line, illustrating the very high resolution of the technique. The first dimensional separation (right to left of figure) was achieved using immobilised pH gradient electrophoresis of 4.0 to 10.0 units. The second dimension (top to bottom of figure) was SDS-PAGE using a 11%–14% acrylamide gradient, allowing separation in the molecular weight range 10–250 kDa. Proteins were visualised by silver staining. Arrows show proteins of known identity.

1992; Celis *et al.*, 1993; Garrels and Franza, 1989; VanBogelen *et al.*, 1992). Current protocols can resolve two to three thousand proteins from a complex sample on a single gel (*Figure 1*).

#### 2-D GEL RESOLUTION AND REPRODUCIBILITY

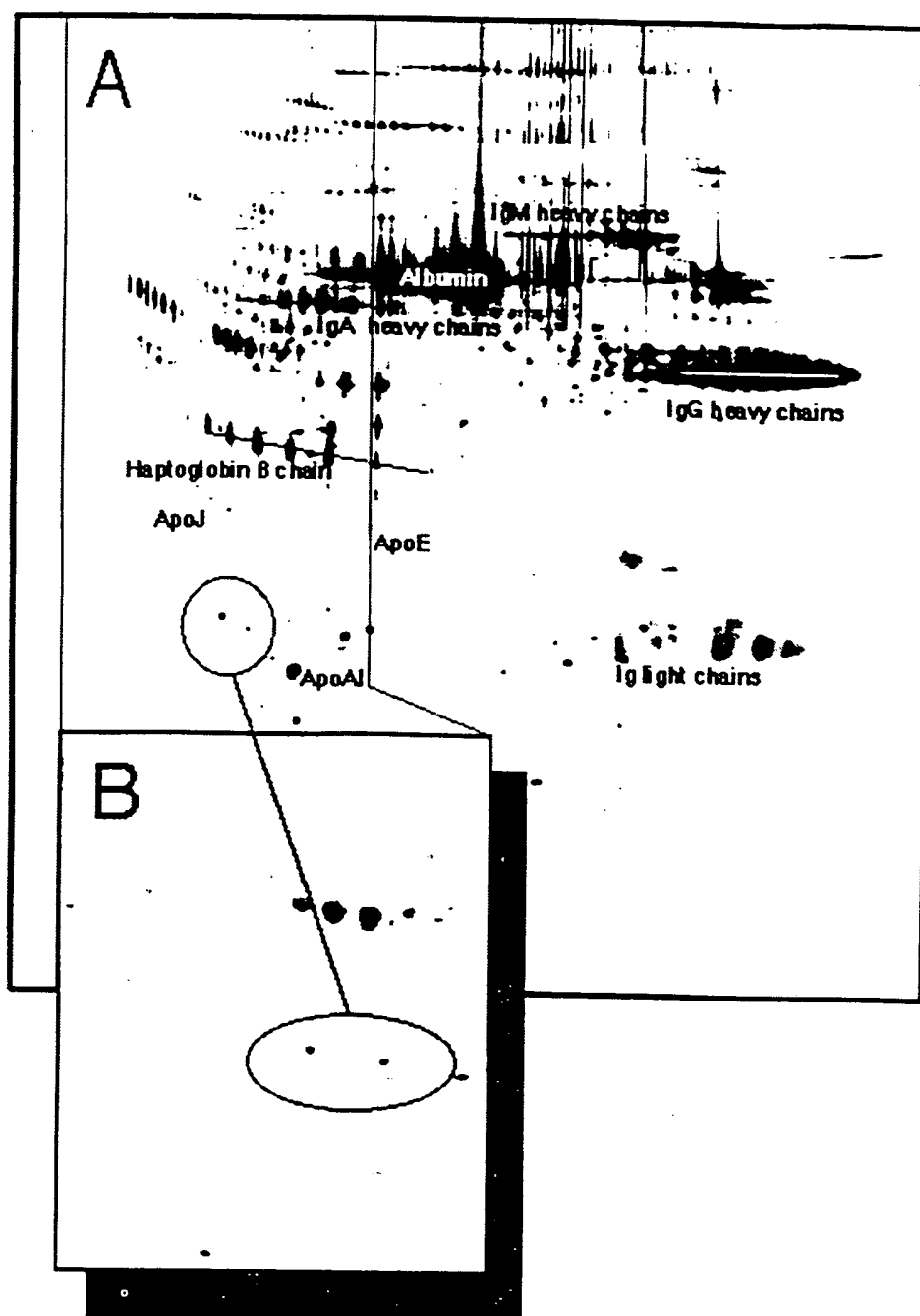
A primary challenge of separating complex mixtures of proteins by 2-D gel electrophoresis has been to achieve high resolution and reproducibility. High resolution ensures that a maximum of protein species are separated, and high reproducibility is

vital to allow comparison of gels from day to day and between research sites. These factors can be difficult to achieve.

Carrier ampholytes are a common means of isoelectric focusing for the first dimension of 2-D electrophoresis. Gels are usually focused to equilibrium to separate proteins in the pI range 4 to 8, and run in a non-equilibrium mode (NEPHGE) to separate proteins of higher pI (7 to 11.5) (O'Farrell, 1975; O'Farrell, Goodman and O'Farrell, 1977). Unfortunately, the use of carrier ampholytes in the isoelectric focusing procedure is susceptible to 'cathode drift', whereby pH gradients established by prefocusing of ampholytes slowly change with time (Righetti and Drysdale, 1973). Carrier ampholyte pH gradients are also distorted by high salt concentration of samples (Bjellqvist *et al.*, 1982), and by high protein load (O'Farrell, 1975). A further limitation is that isoelectric focusing gels, which are cast and subject to electrophoresis in narrow glass tubes, need to be extruded by mechanical means before application to the second dimension – a procedure that potentially distorts the gel. Nevertheless, many of the above shortcomings can be avoided by loading small amounts of  $^{14}\text{C}$  or  $^{35}\text{S}$  radiolabelled samples (Garrels, 1989; Neidhardt *et al.*, 1989; Vandekerckhove *et al.*, 1990). High sensitivity detection is then achieved through use of fluorography or phosphorimaging plates (Bonner and Laskey, 1974; Johnston, Pickett and Barker, 1990; Patterson and Latter, 1993). However, this approach is only practicable for organisms or tissues that can be radiolabelled.

An alternative technique, which is becoming the method of choice for the first dimension separation of proteins, involves isoelectric focusing in immobilized pH gradient (IPG) gels (Bjellqvist *et al.*, 1982; Görg, Postel and Gunther, 1988; Righetti, 1990). Immobilized pH gradients are formed by the covalent coupling of the pH gradient into an acrylamide matrix, creating a gradient that is completely stable with time. IPG gels are usually poured onto a stiff backing film, which is mechanically strong and provides easy gel handling (Ostergren, Eriksson and Bjellqvist, 1988). The major advantages of IPG separations are that they do not suffer from cathodic drift, they allow focusing of basic and very acidic proteins to equilibrium, pH gradients can be precisely tailored (linear, stepwise, sigmoidal), and that separations over a very narrow pH range are possible (0.05 pH units per cm) (Righetti, 1990; Bjellqvist *et al.*, 1982, 1993a; Sinha *et al.*, 1990; Görg *et al.*, 1988; Gelfi *et al.*, 1987; Gunther *et al.*, 1988). However, it is not currently possible to use IPG gels to separate very basic proteins of isoelectric point greater than 10, although this is under development. Narrow pH range separations are useful to address problems of protein co-migration in complex samples, allowing 'zooming in' on regions of a gel (*Figure 2*). IPG gel strips are now commercially available, which begin to address the problems of intra- and inter-lab isoelectric focusing reproducibility.

There are two means of electrophoresis for the second dimension separation of proteins: vertical slab gels and horizontal ultrathin gels (Görg, Postel, and Gunther, 1988). Both are usually SDS-containing gradient gels of approximately 11% to 15% acrylamide, which separate proteins in the molecular mass range of 10 – 150kD. A stacking gel is not usually used with slab gels, but is necessary when using horizontal gel setups (Görg, Postel and Gunther, 1988). Comparisons have shown that there is little or no difference in the reproducibility of electrophoresis using either approach (Corbett *et al.*, 1994a), but commercially available vertical or horizontal precast gels will provide greater reproducibility for occasional users. For slab gel electrophoresis,



**Figure 2.** Two-dimensional gel electrophoresis allows 'zooming in' on areas of interest. Rings highlight 2 proteins common to each gel. (A) Wide pI range two dimensional electrophoresis map of human plasma proteins. First dimension separation was achieved using an immobilised pH gradient of 3.5 to 10.0 units. The second dimension was SDS-PAGE. Actual gel size was 16cm x 20cm, and proteins were visualised with silver staining. (B) Narrow pI range electrophoresis was used to 'zoom in' on a small region of the plasma map. The first dimension used a narrow range immobilised pH gradient of 4.2 to 5.2 units, and second dimension was SDS-PAGE. Micropreparative loading was used, and the gel blotted to PVDF. Proteins were visualised with amido black. Actual blot size was 16cm x 20cm.

the use of piperazine diacryl as a gel crosslinker and the addition of thiosulfate in the catalyst system has been shown to give better resolution and higher sensitivity detection (Hochstrasser and Merrill, 1988; Hochstrasser, Patchornik and Merrill, 1988).

Notwithstanding the advances described above, there is an increasing demand to improve the reproducibility of 2-D electrophoresis to facilitate database construction and proteome studies. Harrington *et al.* (1993) explain that if a gel resolves 4000 protein spots, and there is 99.5% spot matching from gel to gel, this will produce 20 spot errors per gel. This amount of error, which might accumulate with each gel to gel comparison used in database construction, could produce an unacceptable degree of uncertainty in gel databases. To address these issues, partial automation of large 2-D gel separations has been undertaken (Nokihara, Morita and Kuriki, 1992; Harrington *et al.*, 1993). Although results are preliminary, spot to spot positional reproducibility in one study was found to be threefold improved over manual methods (Harrington *et al.*, 1993). It should be noted that small 2-D gel formats (50 × 43 mm) have been almost completely automated (Brewer *et al.*, 1986), although these are not generally used for database studies.

#### MICROPREPARATIVE 2-D GEL ELECTROPHORESIS

With the advent of affordable protein microcharacterisation techniques, including N-terminal microsequencing, amino acid analysis, peptide mass fingerprinting, phosphate analysis and monosaccharide compositional analysis, a new challenge for 2-D electrophoresis has been to maintain high resolution and reproducibility but to provide protein in sufficient quantities for chemical analysis (high nanogram to low microgram quantities of proteins per spot). This becomes difficult to achieve with very complex samples such as whole bacterial cells, as the initial protein load is divided among 2000 to 4000 protein species. Two approaches are used for producing amounts of material that can be chemically characterised. The first method is to run multiple gels, collect and pool the spots of interest, and subject them to concentration (Ji *et al.*, 1994; Walsh *et al.*, 1995; Rasmussen *et al.*, 1992). In this approach, the concentration process must also act as a purification step to remove accumulated electrophoretic contaminants such as glycine. A more elegant approach has been to exploit the high loading capacity of IPG isoelectric focusing. The high loading capacity of immobilised pH gradients was described early (Ek, Bjellqvist and Righetti, 1983), but has only recently been applied to 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). Up to 15 mg of protein can be applied to a single gel, yielding microgram quantities of hundreds of protein species. A further benefit of this approach is that proteins present in low abundance, which may not be visualised by lower protein loads, are more likely to be detected. The use of electrophoretic or chromatographic prefractionation techniques (Hochstrasser *et al.*, 1991a; Harrington *et al.*, 1992), followed by high loading of narrow-range IPG separations (Bjellqvist *et al.*, 1993b) provides a likely solution to studies on proteins present in low abundance.

#### Methods of protein detection

There are many means for detecting proteins from 2-D gels. The method used will be dictated by factors including protein load on gel (analytical or preparative), the purpose of the gel (for protein quantitation or for blotting and chemical characterisation), and the sensitivity required. The most common means of protein detection and their applications are shown in Table 1. Most detection methods have drawbacks, for

Tab  
Det  
Met  
[PS  
rad  
flu  
pho  
[PS  
Silv  
Coc  
R-2  
Coli  
Zinc  
Pon  
am  
Indi  
Stat  
PVI  
of fi  
ex:  
19:  
are  
(St  
pro

**Table 1:** Common stains for 2-D gels or blots and their applications.

Detection Method	Main applications	Unsuitable applications	Sensitivity	References
[ <sup>35</sup> S] Met or [ <sup>14</sup> C] radiolabelling and fluorography or phosphorimaging	Cell lines, cultured organisms	Samples that cannot be labelled	20 ppm of radiolabel in a spot	Garrels and Franza, 1989; Latham, Garrels and Solter, 1993
[ <sup>35</sup> S]thiourea silver	Extremely high sensitivity gel staining	Preparative 2-D: PVDF or NC membranes	0.4 ng protein on spot or band of gel	Wallace and Saluz, 1992a,b
Silver	Very high sensitivity gel staining, can be mono or polychromatic	Preparative 2-D: PVDF or NC membranes	4 ng protein on spot or band of gel	Rabilloud, 1992; Hochstrasser and Merrill, 1988
Coomassie blue R-250	Staining of gels; staining of PVDF membranes before protein sequencing	Staining prior to direct mass determination from PVDF; amino acid analysis on PVDF; detection of some glycoproteins	40 ng protein on band or spot of gel	Strupat <i>et al.</i> , 1994; Gharahdaghi <i>et al.</i> , 1992; Goldberg <i>et al.</i> , 1988; Sanchez <i>et al.</i> , 1992
Colloidal gold	Staining NC membranes, staining PVDF before direct MALDI-TOF	Gels	60 × higher than coomassie	Yamaguchi and Asakawa, 1988; Eckerskorn <i>et al.</i> , 1992; Strupat <i>et al.</i> , 1994
Zinc imidazole	Reverse staining of gels or membranes; may be beneficial in MALDI-TOF of peptides	Where positive image is required	Higher than coomassie	Ortiz <i>et al.</i> , 1992; James <i>et al.</i> , 1993
Ponceau S and amido black	Staining higher protein loads on PVDF, for protein sequencing or amino acid analysis	Staining prior to direct mass determination from PVDF	100 ng protein on band or spot of gel	Sanchez <i>et al.</i> , 1992; Strupat <i>et al.</i> , 1994; Wilkins <i>et al.</i> , 1995
India ink	Staining of membrane-bound proteins; staining PVDF before direct MALDI-TOF	Gel staining; not quantitative from protein to protein	1–10 ng	Li <i>et al.</i> , 1989; Hughes, Mack and Hamparian, 1988; Strupat <i>et al.</i> , 1994
Stains-all	Staining to detect glycoproteins or Ca <sup>2+</sup> binding proteins	General gel staining	100 ng protein on band or spot of gel	Campbell, MacLennan and Jorgensen, 1983; Goldberg <i>et al.</i> , 1988

PVDF = polyvinylidene difluoride, NC = nitrocellulose; MALDI-TOF = matrix assisted laser desorption ionisation time of flight mass spectrometry.

example, some glycoproteins are not stained by coomassie blue (Goldberg *et al.*, 1988), and many organic dyes are unsuitable for protein detection on PVDF if samples are to be used for direct matrix-assisted laser desorption ionisation mass spectrometry (Strupat *et al.*, 1994).

Although most means of protein detection give some indication of the quantities of protein present, in general they cannot be used for global quantitation. This is because

no protein stain is able consistently to detect proteins over a wide range of concentrations, isoelectric points and amino acid compositions, and with a variety of post-translational modifications (Goldberg *et al.*, 1988; Li *et al.*, 1989). Furthermore, there are large differences in staining pattern when identical gels or blots are subjected to different stains, including amido black, imidazole zinc, india ink, ponceau S, colloidal gold, or coomassie blue (Tovey, Ford and Baldo, 1987; Ortiz *et al.*, 1992). The most common means of quantitating large numbers of proteins in a 2-D gel involves the radiolabelling of protein samples prior to electrophoresis, and protein quantitation based on fluorography and image analysis or liquid scintillation counting (Garrels, 1989; Celis and Olsen, 1994). However, proteins which do not contain methionine cannot be detected if only [ $^{35}\text{S}$ ] methionine is used for labelling. Amino acid analysis of protein spots visualised by other techniques presents a likely means of protein quantitation for the future.

#### BLOTTING OF PROTEINS TO MEMBRANES

Electrophoretic blotting of proteins from two-dimensional polyacrylamide gels to membranes presents many options for protein identification and microcharacterisation which are not possible when proteins remain in gels. For example, when proteins are blotted to polyvinylidene difluoride (PVDF) membranes, they can be identified by N-terminal sequencing, amino acid analysis, or immunoblotting, or they may be subjected to endoproteinase digestion, monosaccharide analysis, phosphate analysis, or direct matrix-assisted laser desorption ionisation mass spectrometry (Matsudaira, 1987; Wilkins *et al.*, 1995; Jungblut *et al.*, 1994; Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Weizthandler *et al.*, 1993; Murthy and Iqbal, 1991; Eckerskorn *et al.*, 1992). It is possible to combine some of these procedures on a single protein spot on a PVDF membrane (Packer *et al.*, 1995; Wilkins *et al.*, submitted; Weizthandler *et al.*, 1993). This is useful when minimal amounts of protein are available for analysis. These techniques will be explored in detail later in this review. Notwithstanding the above, there are some disadvantages associated with blotting of proteins to membranes. There is always loss of sample during blotting procedures (Eckerskorn and Lottspeich, 1993), and common protein detection methods are less sensitive or not applicable to membranes (*Table 1*), presenting difficulties for the analysis of low abundance proteins. Detailed discussion of the merits of available membranes and common blotting techniques can be found elsewhere (Eckerskorn and Lottspeich, 1993; Strupat *et al.*, 1994; Patterson, 1994).

#### 2-D gel analysis, documentation, and proteome databases

Following protein electrophoresis and detection, detailed analysis of gel images is undertaken with computer systems. For proteome projects, the aim of this analysis is to catalogue all spots from the 2-D gel in a qualitative and if possible quantitative manner, so as to define the number of proteins present and their levels of expression. Reference gel images, constructed from one or more gels, form the basis of two-dimensional gel databases. These databases also contain protein spot identities and

det  
link  
(Ne  
dat.  
D g  
as g  
Dat

GEL

Afte  
pho  
scar  
Celi  
rese  
or n  
pula  
spot  
spot  
assi  
lists

Tabl

Gel  
ELS  
GEL

MEI  
QUI

TYC

\* Ti  
syst

ity.  
(G  
200  
ima  
ma  
to l  
usi  
Clc  
alt  
19:

details of their post-translational modifications. 2-D gel databases are beginning to be linked to or integrated with comprehensive protein and nucleic acid databases (Neidhardt *et al.*, 1989; Simpson *et al.*, 1992; Appel *et al.*, 1994), and 'organism' databases, containing DNA sequence data, chromosomal map locations, reference 2-D gels and protein functional information for an organism, are becoming established as genome and proteome projects progress (VanBogelen *et al.*, 1992; Yeast Protein Database cited in Garrels *et al.*, 1994).

#### GEL IMAGE ANALYSIS AND REFERENCE GELS

After 2-D electrophoresis and protein visualisation by staining, fluorography or phosphorimaging, images of gels are digitised for computer analysis by an image scanner, laser densitometer, or charge-coupled device (CCD) camera (Garrels, 1989; Celis *et al.*, 1990a; Urwin and Jackson, 1993). All systems digitise gels with a resolution of 100 – 200  $\mu\text{m}$ , and can detect a wide range of densities or shading (256 or more 'grey scales'). Following this, gel images are subjected to a series of manipulations to remove vertical and horizontal streaking and background haze, to detect spot positions and boundaries, and to calculate spot intensity (*Figure 3*). A standard spot (SSP) number, containing vertical and horizontal positional information, is assigned to each detected spot and becomes the protein's reference number. *Table 2* lists some notable software packages which process 2-D gel images.

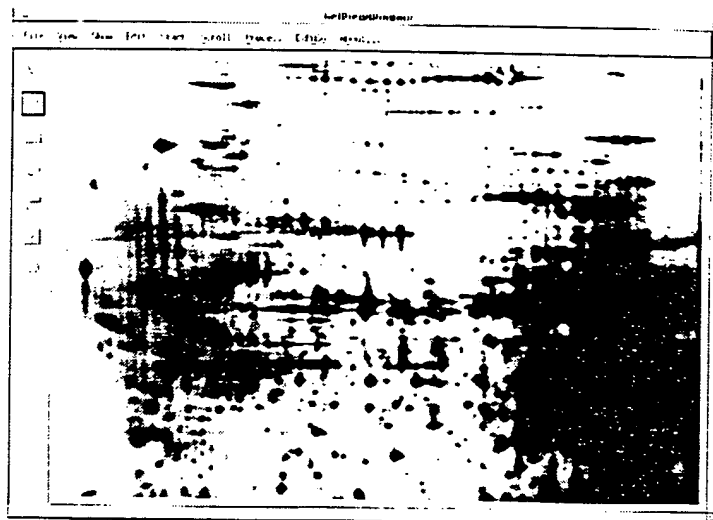
**Table 2:** Some Software Packages for the Analysis of Gel Images.

Gel Image Analysis System	References*
ELSIE 4 & 5	Olsen and Miller, 1988; Wirth <i>et al.</i> , 1991; Wirth <i>et al.</i> , 1993.
GELLAB I & II	Wu, Lemkin and Upton, 1993; Lemkin, Wu and Upton, 1993; Myrick <i>et al.</i> , 1993.
MELANIE I & II	Appel, <i>et al.</i> 1991; Hochstrasser <i>et al.</i> 1991b.
QUEST I & II and PDQUEST	Garrels, 1989; Monardo <i>et al.</i> , 1994; Holt <i>et al.</i> , 1992; Celis <i>et al.</i> , 1990a,b.
TYCHO & KEPLAR	Anderson <i>et al.</i> , 1984; Richardson, Horn and Anderson, 1994.

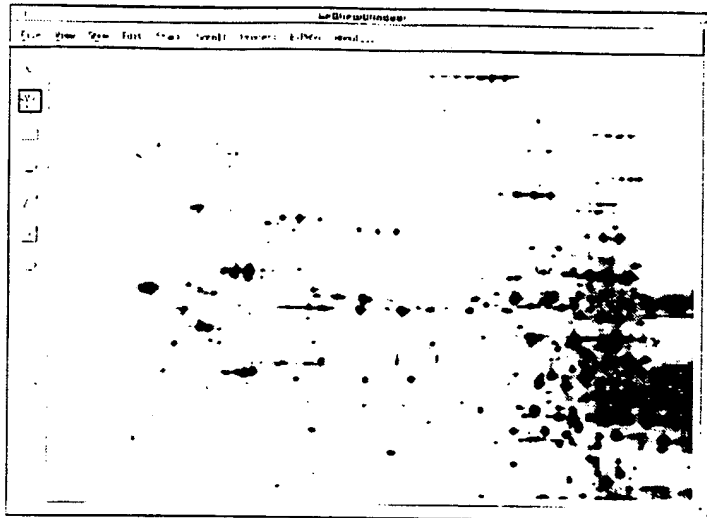
\* These references are not exhaustive; they include some references of use as well as authors of the system.

As there are difficulties in the electrophoresis of samples with 100% reproducibility, reference gel images are often constructed from many gels of the same sample (Garrels and Franza, 1989; Neidhardt *et al.*, 1989). Since this involves the matching of 2000 to 4000 proteins from one gel to another, it presents a considerable challenge to image analysis systems. Matching of gels is usually initiated by an operator, who manually designates approximately 50 or so prominent spots as 'landmarks' on gels to be cross-matched. Proteins which match are then established around landmarks, using computer-based vector algorithms to extend the matching over the entire gel. Close to 100% of spots from complex samples can be matched by these methods, although different degrees of operator intervention may be required (Olsen and Miller, 1988; Lemkin and Lester, 1989; Garrels, 1989; Myrick *et al.*, 1993).

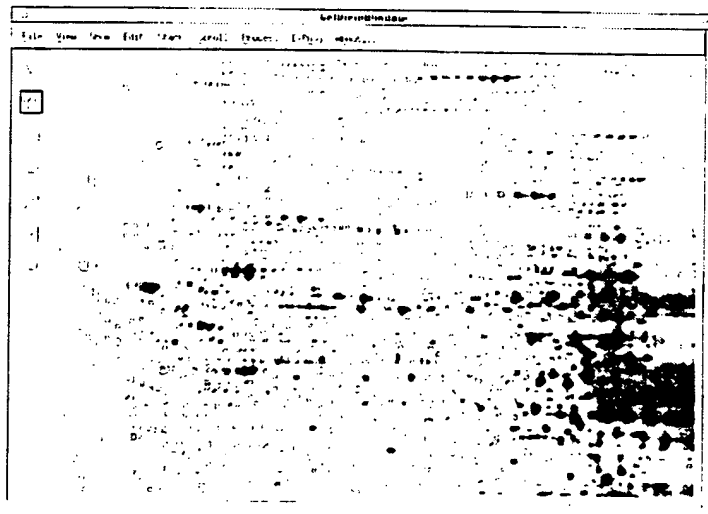
A



B



C



**Figure 3.** Computer processing of gel images. Shown is a wide pI range 2-D separation of human liver proteins, processed by Melanie software (Appel *et al.*, 1991). (A) Original gel image as captured by laser densitometer. (B) Gel image after processing to remove streaking and background. (C) Outline definition of all spots on the gel.

CALCU

Estima  
2-D ge  
during  
are rec  
obtaine  
curves  
MW of  
Bogele  
*et al.*, 19  
to PVE  
*et al.*,  
protein  
amino  
carryin  
positio

SPOT Q

A majo  
separat  
to dete  
positio  
of man  
perform  
minute  
scintill  
by ref  
protein  
*et al.*,  
Garrel  
be not  
Limita  
not act  
only e  
an alte  
Myrici  
1992 ;

Wh  
and m  
protein  
their a  
transf  
regula  
synth  
(Lath



## CALCULATION OF PROTEIN ISOELECTRIC POINT AND MOLECULAR WEIGHT

Estimation of the isoelectric point (pI) and molecular weight (MW) of proteins from 2-D gels provides fundamental parameters for each protein, which are also of use during identification procedures (see following section). The pI and MW of proteins are recorded in 2-D gel databases. Accurate estimations of protein pI and MW can be obtained by using 20 or more known proteins on a reference map to construct standard curves of pI and molecular weight, which are then used to calculate estimated pI and MW of unknown proteins (Neidhardt *et al.*, 1989; Garrels and Franza, 1989; Van-Bogelen, Hutton and Neidhardt, 1990; Anderson and Anderson, 1991; Anderson *et al.*, 1991; Latham *et al.*, 1992). Alternatively, the MW of individual proteins blotted to PVDF can be determined very accurately by direct mass spectrometry (Eckerskorn *et al.*, 1992). Where immobilised pH gradients are used, the focusing position of proteins allows their pI to be measured within 0.15 units of that calculated from the amino acid sequence (Bjellqvist *et al.*, 1993c). It must be noted, however, that proteins carrying post-translational modifications may migrate to unexpected pI or MW positions during electrophoresis (Packer *et al.*, 1995).

## SPOT QUANTITATION AND EXPRESSION ANALYSIS

A major challenge faced in proteome projects is the quantitative analysis of proteins separated by 2-D electrophoresis. The most accurate means of protein quantitation is to determine chemically the amount of each protein present by amino acid compositional analysis. However, the current method of choice for quantitative analysis of many proteins is to radiolabel samples with [<sup>35</sup>S] methionine or <sup>14</sup>C amino acids, perform the 2-D electrophoresis, and measure protein levels in disintegrations per minute (dpm) or units of optical density. Quantitation is achieved either by liquid scintillation counting, or by gel image analysis where spot densities are quantitated by reference to gel calibration strips containing known amounts of radiolabelled protein or against the integrated optical density of all spots visualised (Vandekerckhove *et al.*, 1990; Celis *et al.*, 1990b; Celis and Olsen, 1994; Garrels, 1989; Latham, Garrels and Solter, 1993; Fey *et al.*, 1994). All approaches effectively allow spots to be normalised against the total disintegrations per minute loaded onto the gel. Limitations that remain with radiolabelling methods are that absolute quantitation is not achieved because all proteins have varying amounts of any amino acid, and that only easily labelled samples can be investigated. Quantitative silver staining presents an alternative (Giometti *et al.*, 1991; Harrington *et al.*, 1992; Rodriguez *et al.*, 1993; Myrick *et al.*, 1993), which when undertaken with [<sup>35</sup>S]thiourea (Wallace and Saluz, 1992 a,b) is of extremely high sensitivity.

When protein spots from samples prepared under different conditions are quantitated and matched from gel to gel, it becomes possible to examine changes and patterns in protein expression. Large scale investigation of up- and down-regulation of proteins, their appearance and disappearance, can be undertaken. For example, simian virus 40 transformed human keratinocytes were shown to have 177 up-regulated and 58 down-regulated proteins compared to normal keratinocytes (Celis and Olsen, 1994); detailed synthesis profiles of 1200 proteins have been established in 1 to 4 cell mouse embryos (Latham *et al.*, 1991, 1992); and 4 proteins out of 1971 were found to be markers for

cadmium toxicity in urinary proteins (Myrick *et al.*, 1993). Complex global changes in protein expression as a result of gene disruptions have also been investigated (S. Fey and P. Moser-Larsen, Personal communication). Impressively, large gel sets showing protein expression under different conditions can be globally investigated using statistical methods that find groups of related objects within a set. For example, the REF52 rat cell line database, consisting of 79 gels from 12 experimental groups where each gel contains quantitative data for 1600 cross-matched proteins, has been analysed by cluster analysis (Garrels *et al.*, 1990). This revealed clusters of proteins that, for example, were induced or repressed similarly under simian virus 40 or adenovirus transformation, suggesting a common mechanism. Protein groups that were induced or repressed during culture growth to confluence were also found. It is obvious that the potential for investigation of cellular control mechanisms by these approaches is immense. It is equally clear that investigations of gene expression of this scale are currently technically impossible using nucleic-acid based techniques.

**Table 3:** Some proteome databases and their special features

Proteome database	Special features	References
<i>E. coli</i> gene-protein database	Gel spots linked with GenBank and Kohara clones; quantitative spot measurements under different growth conditions	VanBogelen and Neidhardt, 1991; VanBogelen <i>et al.</i> , 1992
Human heart databases	Identification of disease markers; two separate databases have been established	Baker <i>et al.</i> , 1992 Corbett <i>et al.</i> , 1994b Jungblut <i>et al.</i> , 1994
Human keratinocyte database	Extensive identifications; quantitative spot measurements of transformed cells; identification of disease markers	Celis <i>et al.</i> , 1990a Celis <i>et al.</i> , 1993 Celis and Olsen 1994
Mouse embryo database	Quantitative spot measurements through 1 to 4 cell stage	Latham <i>et al.</i> , 1991 Latham <i>et al.</i> , 1992
Mouse liver database (Argonne Protein Mapping Group)	Documents changes due to exposure to ionizing radiation and toxic chemicals	Giometti, Taylor and Tollaksen, 1992
Rat liver epithelial database	Detailed subcellular fractionation studies	Wirth <i>et al.</i> , 1991 Wirth <i>et al.</i> , 1993
Rat liver database	Extensive studies on regulation of proteins by drugs and toxic agents	Anderson and Anderson, 1991; Anderson <i>et al.</i> , 1992; Richardson, Horn and Anderson, 1994
REF 52 rat cell line database	Accessible via World Wide Web; quantitative spot measurements under different conditions	Garrels and Franza 1989 Boutell <i>et al.</i> , 1994
SWISS-2DPAGE containing human reference maps	Accessible via World Wide Web; completely integrated with SWISS-PROT and SWISS-3DIMAGE	Appel <i>et al.</i> , 1993 Hochstrasser <i>et al.</i> , 1992 Hughes <i>et al.</i> , 1993 Golaz <i>et al.</i> , 1993
Yeast Protein Database (YPD) and Yeast Electrophoretic Protein Database (YEPD)	Completely crossreferenced organism database; YPD has extensive information on over 3500 proteins; YEPD has many identifications	Garrels <i>et al.</i> , 1994

FEATUR  
Protein  
proteins  
informa  
2-D gel  
subcell  
of refer  
should  
Macint  
the are  
annotat  
sequen  
One  
SWISS  
1994;  
feature  
2DPA

**Table 4**  
All three  
expasy

Informa

Annota

Cross-  
Refere  
Datab

Other

## FEATURES OF PROTEOME DATABASES

Proteome projects rely heavily on computer databases to store information about all proteins expressed by an organism. 'Proteome databases' should contain detailed information of proteins already characterised elsewhere, as well as protein data from 2-D gels such as apparent pI and MW, expression level under different conditions, subcellular localisation, and information on post-translational modifications. Images of reference 2-D gels, showing protein SSP numbers and protein identifications, should also be included. Ideally, proteome databases should be accessible with Macintosh or IBM personal computers and easy to use. Some proteome databases and the areas they cover are listed in *Table 3*. Databases range from collections of annotated gels to large databases of images integrated with protein and nucleic acid sequence banks.

One example of an integrated proteome database is the suite of SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE databases (Appel *et al.*, 1993; Appel *et al.*, 1994; Appel, Bairoch and Hochstrasser, 1994; Bairoch and Boeckmann, 1994). The features of these three databases are listed in *Table 4*. SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE are accessible through the World Wide Web

**Table 4:** The SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE suite of crosslinked databases. All three databases are accessible through the World Wide Web, at URL address: <http://expasy.hcuge.ch/>

	SWISS-PROT	SWISS-2DPAGE	SWISS-3DIMAGE
Information	Text entries of sequence data; Citation information; taxonomic data: 38, 303 entries in Release 29	2-D gel images of: human liver, plasma, HepG2, HepG2 secreted proteins, red blood cell, lymphoma, cerebrospinal fluid, macrophage like cell line, erythroleukemia cell, platelet	Collection of 330 3-D images of proteins
Annotations	Protein function; Post translational modifications; Domains; Secondary structure; Quaternary structure; Diseases associated with protein; Sequence conflicts	Gel images where protein is found; How protein identified; Protein pI and MW; protein number; normal and pathological variants	All annotation is available in SWISS- PROT
Cross- Referenced Databases	SWISS-2DPAGE SWISS-3DIMAGE EMBL; PIR; PDB; OMIM; PROSITE; Medline; Flybase; GCRDb; MaizeDB; WormPep; DictyDB	SWISS-PROT and all other databases accessible through SWISS-PROT	SWISS-PROT and all other databases accessible through SWISS-PROT
Other Features	Navigation to other SWISS databases achieved by selecting entries with computer mouse	Gel images show position of identified proteins, or region of gel where protein should appear	Mono and stereo images available; Images can be transferred to local computer image viewing programs

(Berners-Lee *et al.*, 1992), allowing any computer connected to the internet to access the stored information and images. Navigation within and between the three databases is seamless, as all potential crosslinks are highlighted as hypertext on the display and can be selected with a computer mouse. From these databases, detailed information about a protein, including amino acid sequence and known post-translational modifications, can be obtained, the precise protein spot it corresponds to on a reference gel image can be viewed if known, and the 3-D structure of the molecule can be seen if available. References to nucleic acid and other databases are also given to provide access to information stored elsewhere.

'Organism' databases, containing detailed protein and nucleic acid information about a species, are becoming common as genome and proteome projects progress. These differ from nucleic acid or protein sequence databases like GenBank or SWISS-PROT because they are image based, and contain information about chromosomal map positions, transcription of genes, and protein expression patterns. The *Escherichia coli* gene-protein database (VanBogelen, Hutton and Neidhardt, 1990; VanBogelen and Neidhardt, 1991; VanBogelen *et al.*, 1992), known as the ECO2DBASE, is one example. It contains gene and protein names, 2-D gel spot information (including pI and MW estimates, and spot identification), genetic information (GenBank or EMBL codes, chromosomal location, location on Kohara clones (Kohara, Akiyama, and Isono, 1987), transcription direction of genes), and protein regulatory information (level of protein expression under different growth regimes, member of regulon or stimulon). All entries in the ECO2DBASE are also cross-referenced to the SWISS-PROT database (Bairoch and Boeckmann, 1994). It is anticipated that organism databases will soon become a standard means of storing all available information about a particular species. However there is currently no consistent manner in which organism databases are assembled, which may hamper comparisons in the future.

### Identification and characterisation of proteins from 2-D gels

The number of proteins identified on a 2-D reference map determines its usefulness as a research and reference tool. As most reference maps have only a small proportion of proteins identified, a major aim of current proteome projects is to screen many proteins from 2-D maps, in order to define them as 'known' in current nucleic acid and protein databases, or as 'unknown'. Protein identification assists in confirmation of DNA open reading frames, and provides focus for DNA sequencing projects and protein characterisation efforts by pointing to proteins that are novel. Since there may be 3000–4000 proteins from a single 2-D map that require identification, the challenge in protein screening is to identify proteins quickly, with a minimum of cost and effort.

Traditionally, proteins from 2-D gels have been identified by techniques such as immunoblotting, N-terminal microsequencing, internal peptide sequencing, comigration of unknown proteins with known proteins, or by overexpression of homologous genes of interest in the organism under study (Matsudaira, 1987; Rosenfeld *et al.*, 1992; VanBogelen *et al.*, 1992; Celis *et al.*, 1993; Honore *et al.*, 1993; Garrels *et al.*, 1994). Whilst these techniques are powerful identification tools, they are too expensive or time and labour intensive to use in mass screening programs. A hierarchical approach to mass protein identification has been recently suggested as an

Table  
Rapid  
exper

Order

1

2

3

4

5

6

7

8

9

altern  
use c  
mass  
slow  
the co  
of th  
macl  
cons  
techn  
ident

PROT

Ther  
identThis  
to idThe  
radio*al.*,  
chro

1988

1995

phor

radio

**Table 5:** Hierarchical analysis for mass screening of 2-D separated proteins blotted to membranes. Rapid and inexpensive techniques are used as a first step in protein identification, and slower more expensive techniques are then used if necessary. Table modified from Wasinger *et al.* (1995).

Order	Identification technique	References
1	Amino acid analysis	Jungblut <i>et al.</i> , 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Jungblut <i>et al.</i> , 1994; Wilkins <i>et al.</i> , 1995
2	Amino acid analysis with N-terminal sequence tag	Wilkins <i>et al.</i> , submitted
3	Peptide-mass fingerprinting	Henzel <i>et al.</i> , 1993; Pappin, Hojrup and Bleasby, 1993; James <i>et al.</i> , 1993; Mann, Hojrup and Roepstorff, 1993; Yates <i>et al.</i> , 1993; Mortz <i>et al.</i> , 1994; Sutton <i>et al.</i> , 1995
4	Combination of amino acid analysis and peptide mass fingerprinting	Cordwell <i>et al.</i> , 1995; Wasinger <i>et al.</i> , 1995;
5	Mass spectrometry sequence tag	Mann and Wilm, 1994
6	Extensive N-terminal Edman microsequencing	Matsudaira, 1987
7	Internal peptide Edman microsequencing	Rosenfeld <i>et al.</i> , 1992; Hellman <i>et al.</i> , 1995;
8	Microsequencing by mass spectrometry (electrospray ionisation, post-source decay MALDI-TOF)	Johnson and Walsh, 1992
9	Ladder sequencing	Bartlett-Jones <i>et al.</i> , 1994

alternative to traditional approaches (Table 5; Wasinger *et al.*, 1995). This involves the use of rapid and cheap identification tools such as amino acid analysis and peptide mass fingerprinting as first steps in protein identification, followed by the use of slower, more expensive and time consuming identification procedures if necessary. In the construction of this hierarchy the analysis time, cost per sample and the complexity of the data created has been considered, as whilst some techniques require little machine time per sample, the analysis of data can be quite involved and time consuming. Amino acid analysis and peptide mass-fingerprinting based identification techniques in the hierarchy are discussed in detail below. For review of other protein identification techniques in Table 5, see Patterson (1994) and Mann (1995).

#### PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION

There has been a revival of interest in the use of amino acid composition for identification of proteins from 2-D gels after early work by Eckerskorn *et al.* (1988). This technique uses a protein's idiosyncratic amino acid composition profile in order to identify it by comparison with theoretical compositions of proteins in databases. The amino acid composition of proteins can be determined by differential metabolic radiolabelling and quantitative autoradiography after 2-D electrophoresis (Garrels *et al.*, 1994; Frey *et al.*, 1994), or by acid hydrolysis of membrane-blotted proteins and chromatographic analysis of the resulting amino acid mixture (Eckerskorn *et al.*, 1988; Tous *et al.*, 1989; Gharahdaghi *et al.*, 1992; Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). As differential metabolic labelling experiments require X-ray film or phosphor-image plate exposures of up to 140 days, and can only be undertaken with easily radiolabelled samples, the technique is not as rapid or widely applicable as chromato-

## Spot ECOLI-B1M

=====

## Composition:

Asx: 13.2 Glx: 10.4 Ser: 5.7 His: 0.7  
 Gly: 5.4 Thr: 3.8 Ala: 6.7 Pro: 7.9  
 Tyr: 1.3 Arg: 5.0 Val: 8.0 Met: 0.3  
 Ile: 5.9 Leu: 8.0 Phe: 13.3 Lys: 4.4

pI estimate: 6.89 Range searched: ( 6.64, 7.14)

Mw estimate: 16800 Range searched: (13440, 20160)

Closest SWISS-PROT entries for the species ECOLI matched by AA composition:

Rank	Score	Protein	pI	Mw	Description
1	24	<b>PYRI_ECOLI</b>	<b>6.84</b>	<b>16989</b>	<b>ASPARTATE CARBAMOYLTRANSFERASE</b>
2	39	COAA_ECOLI	6.32	36359	PANTOTHENATE KINASE (EC 2.7.1.33)
3	40	META_ECOLI	5.06	35713	HOMOSERINE O-SUCCINYLTRANSFERASE
4	42	CADC_ECOLI	5.52	57812	TRANSCRIPTIONAL ACTIVATOR CADC.
5	43	HLYC_ECOLI	8.58	19769	HEMOLYSIN C, PLASMID.

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified range:

Rank	Score	Protein	pI	Mw	Description
1	24	<b>PYRI_ECOLI</b>	<b>6.84</b>	<b>16989</b>	<b>ASPARTATE CARBAMOYLTRANSFERASE</b>
2	102	TRJ8_ECOLI	6.73	17921	TRAJ PROTEIN.
3	112	YAJG_ECOLI	6.79	19028	HYPOTHETICAL LIPOPROTEIN YAJG.
4	140	YFJB_ECOLI	6.83	14945	HYPOTHETICAL 14.9 KD PROTEIN IN GRPE
5	142	YAHA_ECOLI	7.06	14726	HYPOTHETICAL PROTEIN IN BETT 3'REGION

**Figure 4.** Computer printout from ExPASy server where the empirical amino acid composition, estimated pI and MW of a protein from a 2-D reference map of *E. coli* were matched against all entries in SWISS-PROT for *E. coli*. The correct identification, aspartate carbamoyltransferase, is shown in bold. Low scores indicate a good match. Note how matching within a defined pI and MW range (lower set of proteins) has greatly increased the score difference between the first and second ranking proteins. This score difference gives high confidence in the identification, and is only observed where the top ranking protein is the correct identification (Wilkins *et al.*, 1995).

graphy-based analysis. Proteins blotted to PVDF membranes can be hydrolysed in 1 h at 155°C, amino acids extracted in a single brief step, and each sample automatically derivatised and separated by chromatography in under 40 minutes (Wilkins *et al.*, 1995; Ou *et al.*, 1995). In this manner, one operator can routinely analyse 100 proteins per week on one HPLC unit. This technology lends itself to automation, and it is anticipated that instruments with even greater sample throughput will be developed. When proteins have been prepared by micropreparative 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b), blotted to a PVDF membrane and stained with amido black, any visible protein spot is of sufficient quantity for amino acid analysis (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995).

After the amino acid composition of a protein has been determined, computer programs are used to match it against the calculated compositions of proteins in databases (Eckerskorn *et al.*, 1988; Sibbald, Sommerfeldt and Argos, 1991; Jungblut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995). Matching is usually done with only 15 or 16 amino acids, as cysteine and

Spot  
=====  
Compo  
Asx:  
Gly:  
Tyr:  
Ile:  
pI e:  
Mw e:  
Clos  
rang  
Rank  
=====  
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
**Figure 5**  
same sam  
acid con  
PROT f  
for those  
large see  
the corr  
protein  
tryptop  
to thei  
The co  
a scor  
restric  
1994;  
*et al.*  
match  
in Fig  
refere  
romyc  
lympl  
*et al.*  
PROTE  
SEQU  
Whe

## Spot ECOLI-ACJ

=====

## Composition:

Asx: 9.6 Glx: 10.8 Ser: 4.1 His: 2.7  
 Gly: 12.2 Thr: 3.8 Ala: 11.9 Pro: 3.2  
 Tyr: 6.0 Arg: 3.7 Val: 9.5 Met: 0.6  
 Ile: 5.6 Leu: 8.2 Phe: 3.2 Lys: 4.9

pI estimate: 5.99 Range searched: ( 5.74, 6.24)

Mw estimate: 45000 Range searched: (36000, 54000)

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified range:

Rank	Score	Protein	pI	Mw	N-terminal Seq.
1	21	<b>GLYA_ECOLI</b>	6.03	45316	<b>M L K R E</b>
2	32	YJGB_ECOLI	5.86	36502	M S M I K
3	38	GABT_ECOLI	5.78	45774	M S N S K
4	44	YIHS_ECOLI	5.86	48018	M R I K Y
5	45	DHE4_ECOLI	5.98	48581	M D Q T Y
6	46	ARGD_ECOLI	5.79	43765	M A I E Q
7	46	MURB_ECOLI	5.78	37851	M N H S L
8	47	GLMU_ECOLI	5.98	49162	M L N N A
9	47	ACKA_ECOLI	5.85	43290	M S S K L
10	50	YJJN_ECOLI	6.01	37064	M E S R I

**Figure 5.** A PVDF protein spot from an *E. coli* 2-D reference map was sequenced for 4 cycles, and the same sample then subject to amino acid analysis. The N-terminal sequence was M L K R. When the amino acid composition of the spot, as well as estimated pI and MW, were matched against all entries in SWISS-PROT for *E. coli*, the above list of best matches was produced. N-terminal sequences are from SWISS-PROT for those entries. The top ranking identification of serine hydroxymethyltransferase (bold) did not show a large score difference between the first and second ranking proteins, giving little confidence in this being the correct protein identification. However, the sequence tag (M L K R) confirmed the identity of the protein as serine hydroxymethyltransferase.

tryptophan are destroyed during hydrolysis, asparagine and glutamine are deamidated to their corresponding acids, and proline is not quantitated in some analysis systems. The computer programs produce a list of best matching proteins, which are ranked by a score that indicates the match quality. Some programs allow matching to be restricted to specific 'windows' of MW and pI (Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995), and to protein database entries for one species (Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). The use of such restrictions increases the power of matching. An example of protein identification by amino acid composition is shown in Figure 4. To date, amino acid composition has been used to identify proteins from reference maps of *Spiroplasma melliferum*, *Mycoplasma genitalium*, *E. coli*, *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, human sera, human heart, human lymphocyte, and mouse brain (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995; Jungblut *et al.*, 1992, 1994; Garrels *et al.*, 1994; Frey *et al.*, 1994).

#### PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION AND N-TERMINAL SEQUENCE TAG

When samples from 2-D gels are not unambiguously identified by amino acid

composition, pI and MW, often the correct identification of that protein is amongst the top rankings of the list (Hobohm, Houthaeve and Sander, 1994; Cordwell *et al.*, 1995; Wilkins *et al.*, 1995). Taking advantage of this observation, we have used the mass spectrometry 'sequence tag' concept (Mann and Wilm, 1994) in developing a combined Edman degradation and amino acid analysis approach to protein identification (Wilkins *et al.*, submitted). This involves the N-terminal sequencing of PVDF-blotted proteins by Edman degradation for 3 or 4 cycles to create a 'sequence tag', following which the same sample is used for amino acid analysis. As only a few amino acids are removed from the protein, its composition is not significantly altered. Furthermore, since only a small amount of protein sequence is required, fast but low repetitive yield Edman degradation cycles can be used. Modifications to current procedures should allow 3 cycles to be completed in 1 h, thereby allowing the screening of 100 or more proteins per week on one automated, multi-cartridge sequenator. Amino acid composition, pI and MW of proteins are matched against databases as described above, and N-terminal sequences of best matching proteins are checked with the 'sequence tag' to confirm the protein identity (Figure 5). This technique will be less useful when proteins are N-terminally blocked, but as only a few N-terminal amino acids are susceptible to the acetyl, formyl, or pyroglutamyl modifications that cause blockage, this may itself provide useful information for sequence tag identification. A strength of N-terminal sequence tag and amino acid composition protein identification is that data generated are quickly and easily interpreted.

#### PROTEIN IDENTIFICATION BY PEPTIDE MASS FINGERPRINTING

Techniques for the identification of proteins by peptide mass fingerprinting have recently been described (Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Mortz *et al.*, 1994; Sutton *et al.*, 1995). This involves the generation of peptides from proteins using residue-specific enzymes, the determination of peptide masses, and the matching of these masses against theoretical peptide libraries generated from protein sequence databases. As proteins have different amino acid sequences, their peptides should produce characteristic 'fingerprints'.

The first step of peptide mass fingerprinting is protein digestion. Proteins within the gel matrix or bound to PVDF can be enzymatically digested *in situ*, although *in situ* digests are reported to produce more enzyme autodigestion products, which complicate subsequent peptide mass analysis (James *et al.*, 1993; Rasmussen *et al.*, 1994; Mortz *et al.*, 1994). The enzyme of choice for digestion is currently trypsin (of modified sequencing grade), but other enzymes (Lys-C or *S. aureus* V8 protease) have also been used (Pappin, Hojrup and Bleasby, 1993). To maximise the number of peptides obtained, it is desirable for protein samples to be reduced and alkylated prior to digestion (Mortz *et al.*, 1994; Henzel *et al.*, 1993). This ensures that all disulfide bonds of the protein are broken, and produces protein conformations that are more amenable to digestion. Surprisingly, chemical digestion methods such as cyanogen bromide (methionine specific), formic acid (aspartic acid specific), and 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine (tryptophan specific) have not been explored as means of peptide production for mass fingerprinting, even though they are rapid and may circumvent some problems associated with enzyme digestions



(Nikodem and Fresco, 1979; Crimmins *et al.*, 1990; Vanfleteren *et al.*, 1992).

After proteins are digested, peptide masses are determined by mass spectrometry. Direct analysis of peptide mixtures can be achieved by electrospray ionisation mass spectrometry, plasma desorption mass spectrometry, or matrix assisted laser desorption ionization (MALDI) mass spectrometry techniques. MALDI is preferable because of its higher sensitivity and greater tolerance to contaminating substances from 2-D gels (James *et al.*, 1993; Mortz *et al.*, 1994; Pappin, Hojrup and Bleasby, 1993). Furthermore, recent modifications to sample preparation methods have largely solved early difficulties experienced with the calibration of MALDI spectra (Mortz *et al.*, 1994; Vorm and Mann, 1994; Vorm, Roepstorff and Mann, 1994). The high sensitivity of mass spectrometry allows a small fraction of a digest of a 1 µg protein spot to be used for analysis, and analysis itself is complete in a few minutes.

A major challenge associated with peptide mass fingerprinting is data interpretation prior to computer matching against libraries of theoretical peptide digests. Spectra must be examined carefully to determine which peaks represent peptide masses of interest, as there are often enzyme autodigestion products and contaminating substances present (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). Furthermore, if protein alkylation and reduction has not been undertaken prior to protein digestion, peptide sequence coverage may be poor (40% to 70%), with some masses present representing disulfide bonded peptides originally present in the protein (Mortz *et al.*, 1994). For eukaryotes, a serious issue is the alteration of peptide masses by the presence of post-translational modifications (Table 6). The mass of the unmodified peptide alone can be very difficult to determine. Two artifactual modifications introduced by electrophoresis, an acrylamide adduct to cysteine and the oxidation of methionine, are also known to alter peptide masses (le Maire *et al.*, 1993; Hess *et al.*, 1993).

**Table 6:** Masses of some common post-translational modifications. Peptides carrying post-translational modifications complicate data analysis for peptide mass fingerprinting protein identification. This is especially so for protein glycosylation, which involves many different combinations of the hexosamines, hexoses, deoxyhexoses, and sialic acid.

Post-translational modification	Mass change
Acetylation	+ 42.04
*Acrylamide adduct to cysteine	+71.00
Carboxylation of Asp or Glu	+ 44.01
Deamidation of Asn or Gln	+ 0.98
Disulfide bond formation	- 2.02
Deoxyhexoses (Fuc)	146.14
Formylation	+ 28.01
Hexosamines (GlcN, GalN)	+ 161.16
Hexoses (Glc, Gal, Man)	+ 162.14
Hydroxylation	+ 16.00
N-acetylhexosamines (GlcNAc, GalNAc)	+ 203.19
*Oxidation of Met	+ 16.00
Phosphorylation	+ 79.98
Pyroglutamic acid formed from Gln	-17.03
Sialic acid (NeuNAc)	+ 291.26
Sulfation	+ 80.06

Table modified from Finnigan LASERMAT application data sheet 5.

Asterisk \* shows modifications that can arise artifactually from the 2-D electrophoresis process.

A number of computer programs are available for matching peptide masses against databases (reviewed in Cottrell, 1994). Matching is usually undertaken in an interactive manner, whereby peaks of mass 500–3000 Da are selected and matched under various search parameters including MW of protein, mass accuracy of peptides, and number of missed enzyme cleavages allowed (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). The correct protein identity is the protein which has the most peptide masses in common with the unknown sample. Identities have been established with as few as three peptides, but unambiguous identification is thought to require a mass spectrometric map covering most peptides of the protein (Mortz *et al.*, 1994; Yates *et al.*, 1993). To date, peptide mass fingerprinting of proteins has been undertaken from the human myocardial protein and keratinocyte maps, from an *E. coli* 2-D gel, and from reference maps of *Spiroplasma melliferum* and *Mycoplasma genitalium* (Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Henzel *et al.*, 1993; Cordwell *et al.*, 1995; Wasinger *et al.*, 1995), although the technique is most powerful when used in combination with another protein identification technique (Rasmussen *et al.*, 1994; Cordwell *et al.*, 1995).

#### MASS SPECTROMETRY SEQUENCE TAGGING

An extension of peptide mass fingerprinting has recently been described, called peptide sequence tagging (Mann and Wilm, 1994; Mann, 1995). This uses tandem mass spectrometry (MS/MS) to initially determine the mass of peptides, then subject them to fragmentation by collision with a gas, and finally determine the mass of fragments. The resulting spectra gives information about a peptide's amino acid sequence. The fragmentation masses of peptides can rarely be used to assign a complete sequence, but it usually allows a short 'sequence tag' of 2 or 3 amino acids to be determined. This sequence tag and the original peptide mass is matched by computer against a database, providing a likely identity of the peptide and the protein it came from. The major drawback for this technique as a mass screening tool is the complexity of the mass data generated and the high level of expertise required for its interpretation. Nevertheless, it represents a useful new protein identification method which greatly increases the power of peptide mass fingerprinting protein identification.

#### Cross-species protein identification

Protein sequence databases continue to grow at a rapid rate, yet it is not widely appreciated that close to 90% of all information contained in current protein databases comes from only 10 species (A. Bairoch, Pers. Comm.). Fortunately, this information can be used to study proteomes of organisms that are poorly defined at the molecular level, via 2-D electrophoresis and 'cross-species' protein identification (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). This approach allows proteins from reference maps of many different species to be identified without the need for the corresponding genes to be cloned and sequenced. This is particularly true for 'housekeeping' proteins, such as enzymes involved in glycolysis, DNA manipulation and protein manufacture, which are highly conserved across species boundaries. Proteins that cannot be identified across species boundaries can then become the focus of further protein characterisation and DNA sequencing efforts.

A)

Pr  
==As  
Gly  
Ty  
IlepI  
Mw

The

Ran  
==

B)

Rea

Sca

195

130

100

70

No

1

2

3

4

5

6

7

8

9

10

**Figure**  
and try  
be con  
identit  
and He  
all ent  
protein  
progra  
match  
apolip

## A)

## Protein APA1\_HUMAN

=====

Asx: 8.4 Clx: 19.3 Ser: 6.3 His: 1.3  
 Gly: 4.2 Thr: 4.3 Ala: 8.0 Pro: 4.2  
 Tyr: 2.9 Arg: 6.7 Val: 5.5 Met: 1.3  
 Ile: 0.0 Leu: 15.5 Phe: 2.5 Lys: 8.8

pI Range: no range specified

Mw Range: no range specified

The closest SWISS-PROT entries are:

Rank	Score	Protein	(pI	Mw)	Description
1	0	APA1_HUMAN	5.27	28078	APOLIPOPROTEIN A-I.
2	4	APA1_MACFA	5.43	28005	APOLIPOPROTEIN A-I.
3	12	APA1_RABIT	5.15	27836	APOLIPOPROTEIN A-I.
4	14	APA1_BOVIN	5.36	27549	APOLIPOPROTEIN A-I.
5	14	APA1_CANFA	5.10	27467	APOLIPOPROTEIN A-I.
6	18	APA1_MOUSE	5.42	27922	APOLIPOPROTEIN A-I.
7	26	APA1_PIG	5.19	27598	APOLIPOPROTEIN A-I.
8	27	APA1_CHICK	5.26	27966	APOLIPOPROTEIN A-I.
9	37	DYNA_CHICK	5.44	117742	DYNACTIN, 117 KD ISOFORM.
10	39	APA4_HUMAN	5.18	43374	APOLIPOPROTEIN A-IV.

## B)

Reagent: Trypsin MW filter: 10%

Scan using fragment mws of:

1953	1933	1731	1613	1401	1387
1301	1283	1252	1235	1231	1215
1031	896	873	831	813	781
732	704				

No. of database entries scanned = 72018

1	. APA1_HUMAN	APOLIPOPROTEIN A-I (APO-AI). - HOMO SAPIENS
2	. APA1_MACFA	APOLIPOPROTEIN A-I (APO-AI). - MACACA FASCICULARIS
3	. APA1_PAPHA	APOLIPOPROTEIN A-I (APO-AI). - PAPIO HAMADRYAS
4	. B41845	orf B - Treponema denticola
5	. APA1_CANFA	APOLIPOPROTEIN A-I (APO-AI). - CANIS FAMILIARIS (DOG).
6	. S30947	hypothetical protein 1 - Azotobacter vinelandii
7	. HS2C_PEA	CHLOROPLAST HEAT SHOCK PROTEIN PRECURSOR. - PISUM SATIVU
8	. S20724	Tropomyosin - African clawed frog
9	. HIVVI354	HIVVI354 premature term. at 793 - Human immunodeficiency
10	. TRJ2_ECOLI	TRAJ PROTEIN. - ESCHERICHIA COLI.

**Figure 6.** Theoretical cross-species matching of human apolipoprotein A-I by amino acid composition and tryptic peptides. When an unknown protein is analysed, best ranking proteins from both techniques can be compared. If the same protein type is observed in both lists, there is high confidence in this being the identity of the unknown molecule (Cordwell *et al.*, 1995). (A) Output of ExPASy server (Appel, Bairoch and Hochstrasser, 1994) where the true amino acid composition of apolipoprotein A-I was matched against all entries in the SWISS-PROT database, without pI or MW windows. Seven of the top 10 matching proteins were apolipoprotein A-I of different species. (B) Output of MOWSE peptide mass fingerprinting program (Pappin, Hojrup and Bleasby, 1993) where true tryptic peptides of human apolipoprotein A-I were matched against the OWL database, using MW window of 10%. Four of the top ten matching proteins were apolipoprotein A-I from different species.

Rapid cross-species identification of proteins from 2-D reference maps can be undertaken with amino acid composition or peptide mass fingerprinting methods (Figure 6), but these techniques alone may not identify proteins unambiguously when phylogenetic cross-species distances are great or analysis data is of poor quality (Yates *et al.*, 1993; Shaw, 1993; Cordwell *et al.*, 1995). However, very high confidence in protein identities can be achieved when lists of best-matching proteins generated by both techniques are compared (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). The correct identification is found when the same protein is ranked highly in lists of best matches generated by both techniques. This method has allowed approximately 120 proteins from the reference map of the mollicute *Spiroplasma melliferum*, representing approximately one quarter of the proteome, to be confidently identified by reference to protein information from other species (S. Cordwell, Personal Communication). When cross-species protein identification is to be undertaken, it should be noted that the molecular weight of a protein type across species is usually highly conserved, but that protein pI can vary by more than 2 units (Cordwell *et al.*, 1995). Accurate molecular weight determination by direct mass spectrometry of proteins blotted to PVDF (Eckerskorn *et al.*, 1992) should therefore be a useful additional parameter for cross-species protein identification.

#### CHARACTERISATION OF POST-TRANSLATIONAL MODIFICATIONS

Many proteins are modified after translation. Such post-translational modifications, including glycosylation, phosphorylation, and sulfation (see Table 6), are usually necessary for protein function or stability. Some abnormal modifications are associated with disease (Duthel and Revol, 1993; Ghosh *et al.*, 1993; Yamashita *et al.*, 1993). In proteome studies, post-translational modifications can be examined on all proteins present, or on individual spots. Studies on all proteins provide an indication of which proteins may carry a certain type of modification. For example, 2-D gel analysis of cell cultures grown in the presence of [<sup>3</sup>H] mannose or [<sup>32</sup>P] phosphate gives an indication of which proteins carry glycans containing mannose, and which proteins are phosphorylated (Garrels and Franza, 1989). Lectin binding studies of 2-D gels blotted to PVDF or nitrocellulose provide information on the saccharides, if any, that are carried by proteins present (Gravel *et al.*, 1994).

When individual proteins of interest carrying post-translational modifications have been found, micropreparative 2-D electrophoresis can be used to purify them in microgram quantities (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). If protein isoforms of similar MW and pI are to be studied, focusing with narrow range pI gradients (1 pH unit) can provide greater separation and resolution. After electrophoresis, the type and degree of protein phosphorylation can be investigated (Murthy and Iqbal, 1991; Gold *et al.*, 1994), monosaccharide composition can be determined (Weitzhandler *et al.*, 1993; Packer *et al.*, 1995), and the structure and exact site of glycoamino acids can be investigated by either Edman degradation based techniques or by mass spectrometry (Pisano *et al.*, 1993; Huberty *et al.*, 1993; Carr, Huddleston and Bean, 1993). With further development of rapid techniques, investigation of phosphorylation and monosaccharides by chromatographic or mass spectrometric means is likely to become a routine step in the characterisation of post-translational modifications of proteins from reference maps.

The s  
Many  
review  
Adva  
initia  
each  
prote  
genom  
size o  
compi  
this, 2  
plasm  
Wasin  
maps  
specie  
and si

**Table 7**  
PROTE  
referen  
from B  
1996.

Species

*Mycopla*  
*Escheri*  
*Saccha*  
*Dictyos*  
*Arabido*  
*Caenor*  
*Homo s*

The  
under  
becau  
hundr  
estima  
to tiss  
protei  
electro  
ism c  
accele  
are ur  
post-t  
differ  
useful

## The status of proteome projects

Many technical aspects of proteome research have already been discussed in this review, but an overview of the status of proteome projects has not yet been presented. Advances in proteome projects will initially rely on progress in genome sequencing initiatives, to enable an identity, amino acid sequence, or function to be assigned to each protein spot. Table 7 shows genome size, proteome size, and the number of proteins already defined for a number of model organisms. This indicates that whilst genome sequencing programs for *E. coli* and *S. cerevisiae* are advanced, the massive size of some other genomes (and especially the human genome) means that their complete nucleotide sequences are unlikely to be available for many years. Because of this, 2-D reference maps and proteome projects of single cell organisms like *Mycoplasma sp.*, *E. coli* and *S. cerevisiae* will be the most detailed (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Vanbogelen *et al.*, 1992; Garrels *et al.*, 1994), and complete maps of other organisms will take longer to construct. However, the use of cross-species protein identification techniques will allow proteomes of many prokaryotes and simple eukaryotes to be partially defined in reference to *E. coli* and *S. cerevisiae*.

**Table 7:** Estimated genome size, estimated proteome size, number of protein sequences in SWISS-PROT Release 31 (March, 1995), and approximate number of proteins of known identity on 2-D reference maps for some model organisms. Genome size data from Smith (1994), and total protein data from Bird (1995). Genome sequencing projects of *E. coli* and *S. cerevisiae* will probably be complete in 1996.

Species Name	Haploid genome size (million bp)	Estimated proteome size (total proteins)	Protein entries in SWISS PROT	Proteins annotated on 2-D Maps
<i>Mycoplasma species</i>	0.6–0.8	400–600	100	> 100
<i>Escherichia coli</i>	4.8	4000	3170	> 300
<i>Saccharomyces cerevisiae</i>	13.5	6000	3160	> 100
<i>Dictyostelium discoideum</i>	70	12500	204	–
<i>Arabidopsis thaliana</i>	70	14000	270	–
<i>Caenorhabditis elegans</i>	80	17800	703	–
<i>Homo sapiens</i>	2900	60000–80000	3326	> 1000

The study of vertebrate proteomes and vertebrate development is a phenomenal undertaking in comparison to the investigation of single cell organisms. This is because vast numbers of proteins are developmentally expressed, each body tissue has hundreds of unique proteins, and there are numerous tissue types. However, it is estimated that at least 35% of proteins in vertebrate cells will be conserved from tissue to tissue, constituting the 'housekeeping' proteins (Bird, 1995), with the remainder of proteins constituting a set that are specific to a cell type. Providing that standardised electrophoretic conditions are used, reference maps from many tissues of one organism can be superimposed in gel databases (e.g. Hochstrasser *et al.*, 1992). This accelerates the definition of the 'housekeeping' proteins, as well as sets of proteins that are unique to different tissue types. Such studies may, however, be complicated by post-translational modifications, which can differ on the same gene product in different tissues. Proteins that remain unknown after identification procedures will be useful in providing focus for nucleic acid sequencing initiatives.

## FUTURE DIRECTIONS OF PROTEOME PROJECTS

This review has described recent advances in the area of proteome research. It has illustrated how new developments of older techniques (2-D electrophoresis and amino acid analysis) as well as the applications of new technology (mass spectrometry) have greatly widened the choice of tools the biologist and protein chemist has for the separation, identification and analysis of complex mixtures of proteins. This has made possible the establishment of detailed reference maps for organisms, which are becoming the method of choice for the definition of tissues or whole cells, and the investigation of gene expression therein.

Proteome projects are already impacting on the dogma of molecular biology that DNA sequence constitutes the definition of an organism. For example, the proteomes of different tissues of a single organism are often significantly different. Similarly, cross-species identification of proteins (for example the identification of proteins from *Candida albicans* by comparison with *S. cerevisiae*) can open up studies on organisms that are poorly molecularly defined. As cross-species identification can proceed at a pace orders of magnitude faster than a genome project in terms of defining the gene and protein complement of organisms, the need for the DNA sequencing of genomes will be avoided, and emphasis placed on those found to be novel.

Just as genome sequencing is not an end in itself, neither is an annotated 2-D protein reference map of an organism, nor indeed the identification of proteins in a proteome. So whilst an immediate aim of proteome projects is to screen proteins in reference maps, this will lead to expression studies and characterisation of post-translational modifications. The challenge that then needs to be addressed is the investigation of structure and function of proteins in a proteome. The magnitude of this is illustrated by the fact that over half the open reading frames identified in *S. cerevisiae* chromosome III were initially of no known function (Oliver *et al.*, 1992). Structural and functional studies will be an undertaking just as formidable as genome studies are now and proteome projects are becoming, but will lead to an unimaginably detailed understanding of how living organisms are constructed and how they operate.

## Acknowledgements

MRW is the recipient of an Australian Postgraduate Research Award. AAG, MRW, IHS and K LW acknowledge assistance for proteome projects through Macquarie University Research Grants, the Australian Research Council, the Australian National Health and Medical Research Council, Beckman Instruments and GBC Scientific Equipment. DH acknowledges the financial support of a Montus Foundation Grant and the Swiss National Fund for Scientific Research (Grant # 31-33658.92). We thank colleagues who supplied work that was 'In Press' during the writing of this review.

## References

- ANDERSON, N.L., HOFMANN, J.P., GEMMELL, A. AND TAYLOR, J. (1984). Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clinical Chemistry*, **30**, 2031-2036.

- ANDERSON, N.L. AND ANDERSON, N.G. (1991). A two-dimensional gel database of human plasma proteins. *Electrophoresis*, **12**, 883-906.
- ANDERSON, N.L., ESQUER-BLASCO, R., HOFMANN, J.P. AND ANDERSON, N.G. (1991). A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis*, **12**, 907-930.
- ANDERSON, N.L., COPPLE, D.C., BENDELE, R.A., PROBST, G.S., RICHARDSON, F.C. (1992). Covalent protein modifications and gene expression changes in rodent liver following administration of methapyrilene: a study using two-dimensional electrophoresis. *Fundamental and Applied Toxicology*, **18**, 570-580.
- APPEL, R.D., BAIROCH, A. AND HOCHSTRASSER, D.F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends in Biochemical Sciences*, **19**, 258-260.
- APPEL, R.D., HOCHSTRASSER, D.F., FUNK, M., VARGAS, J.R., PELIGRINI, C., MULLER, A.F. AND SCHERRER, J.R. (1991). The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis*, **12**, 722-735.
- APPEL, R.D., SANCHEZ, J-C., BAIROCH, A., GOLAZ, O., MIU, M., VARGAS, J.R. AND HOCHSTRASSER, D.F. (1993). SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis*, **14**, 1323-1328.
- APPEL, R.D., SANCHEZ, J-C., BAIROCH, A., GOLAZ, O., RAVIER, F., PASQUALI, C., HUGHES, G. AND HOCHSTRASSER, D.F. (1994). The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucleic Acids Research*, **22**, 3581-3582.
- BAIROCH, A. AND BOECKMANN, B. (1994). The SWISS-PROT protein sequence databank: current status. *Nucleic Acids Research*, **22**, 3578-3580.
- BAKER, C.S., CORBETT, J.M., MAY, A.J., YACIOUB, M.H. AND DUNN, M.J. (1992). A human myocardial two-dimensional electrophoresis database: protein characterisation by microsequencing and immunoblotting. *Electrophoresis*, **13**, 723-726.
- BARTLET-JONES, M., JEFFERY, W.A., HANSEN, H.F. AND PAPPIN, D.J.C. (1994). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Communications in Mass Spectrometry*, **8**, 737-742.
- BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. AND STRAUSS, M. (1993). Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272-4280.
- BERNERS-LEE, T.J., CAILIAU, R., GROFF, J.F. AND POLLERMAN, B. (1992). *Electronic Networking: Research, Applications, and Policy*, **2**, 52-58.
- BIRD, A.P. (1995). Gene number, noise reduction and biological complexity. *Trends in Genetics*, **11**, 94-100.
- BJELLQVIST, B., EK, K., RIGHETTI, P.G., GIANAZZA, E., GORG, A., WESTERMEIER, R. AND POSTEL, W. (1982). Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods*, **6**, 317-339.
- BJELLQVIST, B., PASQUALI, C., RAVIER, F., SANCHEZ, J-C. AND HOCHSTRASSER, D.F. (1993a). A nonlinear wide-range immobilized pH gradient for two-dimensional electrophoresis and its definition in a relevant pH scale. *Electrophoresis*, **14**, 1357-1365.
- BJELLQVIST, B., SANCHEZ, J-C., PASQUALI, C., RAVIER, F., PAQUET, N., FRUTIGER, S., HUGHES, G.J. AND HOCHSTRASSER, D.F. (1993b). Micropreparative 2-D electrophoresis allowing the separation of milligram amounts of proteins. *Electrophoresis*, **14**, 1375-1378.
- BJELLQVIST, B., HUGHES, G., PASQUALI, C., PAQUET, N., RAVIER, F., SANCHEZ, J-C., FRUTIGER, S. AND HOCHSTRASSER, D. (1993c). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, **14**, 1023-1031.
- BONNER, W.M. AND LASKEY, R.A. (1974). A film detection method for tritium-labeled proteins and nucleic acids in polyacrylamide gels. *European Journal of Biochemistry*, **46**, 83-88.
- BOUTELL, T., GARRELS, J.I., FRANZA, B.R., MONARDO, P.J. AND LATTER, G.I. (1994). REF52 on global gel navigator: an internet-accessible two-dimensional gel electrophoresis database. *Electrophoresis*, **15**, 1487-1490.
- BREWER, J., GRUND, E., HAGERLID, P., OLSSON, I. AND LIZANA, J. (1986). In *Electrophoresis '86* (M.J. Dunn, Ed.), pp. 226-229. VCH, Weinheim.

- CARR, S.A., HUDDLESTON, M.J. AND BEAN, M.F. (1993). Selective identification and differentiation of *N*- and *O*-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Science*, **2**, 183-196.
- CAMPBELL, K.P., MACLENNAN, D.H. AND JORGENSEN, A.O. (1983). Staining of the  $\text{Ca}^{2+}$ -binding proteins, calsequestrin, calmodulin, troponin C, and S-100, with the cationic dye 'Stains-all'. *Journal of Biological Chemistry*, **258**, 11267-11273.
- CELIS, J.E., CRUGER, D., KIL, J., DEJGAARD, K., LAURIDSEN, J.B., RATZ, G.P., BASSE, B., CELIS, A., RASMUSSEN, H.H., BAUW, G. AND VANDEKERKHOVE, J. (1990a). A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis*, **11**, 242-254.
- CELIS, J.E., GESSER, B., RASMUSSEN, H.H., MADSEN, P., LEFFERS, H., DEJGAARD, K., HONORE, B., OLSEN, E., RATZ, G., LAURIDSEN, J.B., BASSE, B., MOURIZTEN, S., HELLERUP, M., ANDERSEN, A., WALBUM, E., CELIS, A., BAUW, G., PUYPE, M., VAN DAMME, J. AND VANDEKERKHOVE, J. (1990b). Comprehensive two-dimensional gel protein database offers a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis*, **11**, 898-1071.
- CELIS, J.E., RASMUSSEN, H.H., OLSEN, E., MADSEN, P., LEFFERS, H., HONORE, B., DEJGAARD, K., GROMOV, P., HOFFMANN, H.J., NIELSEN, M., VASSILEV, A., VINTERMYR, O., HAO, J., CELIS, A., BASSE, B., LAURIDSEN, J., RATZ, G.P., ANDERSEN, A.H., WALBUM, E., KJAERGAARD, I., PUYPE, M., VAN DAMME, J. AND VANDEKERKHOVE, J. (1993). The human keratinocyte two-dimensional database: update 1993. *Electrophoresis*, **14**, 1091-1198.
- CELIS, J.E. AND OLSEN, E. (1994). A qualitative and quantitative protein database approach identified individual and groups of functionally related proteins that are differentially regulated in simian virus 40 (SV40) transformed human keratinocytes: an overview of the functional changes associated with the transformed phenotype. *Electrophoresis*, **15**, 309-344.
- CORBETT, J.M., DUNN, M.J., POSCH, A. AND GÖRG, A. (1994a). Positional reproducibility of protein spots in two-dimensional polyacrylamide electrophoresis using immobilized pH gradient isoelectric focusing in the first dimension - an interlaboratory comparison. *Electrophoresis*, **15**, 1205-1211.
- CORBETT, J.M., WHEELER, C.H., BAKER, C.S., YACOB, M.H. AND DUNN, M.J. (1994b). The human myocardial two-dimensional gel protein database: update 1994. *Electrophoresis*, **15**, 1459-1465.
- CORDWELL, S., WILKINS, M.R., CERPA-POLJAK, A., GOOLEY, A.A., DUNCAN, M., WILLIAMS, K.L. AND HUMPHERY-SMITH, I. (1995). Cross-species identification of proteins separated by two-dimensional electrophoresis using MALDI-TOF and amino acid composition. *Electrophoresis*, **15**, 438-443.
- COTTRELL, J.S. (1994). Protein identification by peptide mass fingerprinting. *Peptide Research*, **7**, 115-124.
- CRIMMINS, D.L., MCCOURT, D.W., THOMA, R.S., SCOTT, M.G., MACKE, K. AND SCHWARTZ, B.D. (1990). *In situ* chemical cleavage of proteins immobilized to glass-fiber and polyvinylidene fluoride membranes: cleavage at tryptophan residues with 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine to obtain internal amino acid sequence. *Analytical Biochemistry*, **187**, 27-38.
- DUTHEL, S. AND REVOL, A. (1993). Glycan microheterogeneity of alpha 1-antitrypsin in serum and meconium from normal and cystic fibrosis patients by crossed immunoaffinoelectrophoresis with different lectins (Con A, LCA, WGA). *Clinical and Chemical Acta*, **215**, 173-187.
- ECKERSKORN, C., JUNGBLUT, P., MEWES, W., KLOSE, J. AND LOTTSPEICH, F. (1988). Identification of mouse brain proteins after two-dimensional electrophoresis and electroblotting by microsequence analysis and amino acid composition. *Electrophoresis*, **9**, 830-838.
- ECKERSKORN, C., STRUPAT, K., KARAS, M., HILLENKAMP, F. AND LOTTSPEICH, F. (1992).



- Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis*, **13**, 664–665.
- ECKERSKORN, C. AND LOTTSCHEICH, F. (1993). Structural characterisation of blotting membranes and the influence of membrane parameters for electroblotting and subsequent amino acid sequence analysis of proteins. *Electrophoresis*, **14**, 831–838.
- EK, K., BJELLQVIST, B.J. AND RIGHETTI, P.G. (1983). Preparative isoelectric focusing in immobilized pH gradients. I. General principles and methodology. *Journal of Biochemical and Biophysical Methods*, **8**, 135–155.
- FEY, S.J., CARLSEN, J., MOSE LARSEN, P., JENSEN, U.A., KJELDSSEN, K. AND HAUNSO, S. (1994). Two-dimensional gel electrophoresis as a tool for molecular cardiology. Proceedings of the International Society for Heart Research 'XV European Section Meeting', pp 9–16.
- FREY, J.R., KUHN, L., KETTMAN, J.R. AND LEFKOVITS, I. (1994). The amino acid composition of 350 lymphocyte proteins. *Molecular Immunology*, **31**, 1219–1231.
- GARRELS, J.I. (1989). The QUEST system for quantitative analysis of two-dimensional gels. *Journal of Biological Chemistry*, **264**, 5269–5282.
- GARRELS, J.I. AND FRANZA, B.R. (1989). The REF52 protein database. *Journal of Biological Chemistry*, **264**, 5283–5298.
- GARRELS, J.I., FRANZA, B.R., CHANG, C. AND LATTER, G. (1990). Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis*, **11**, 1114–1130.
- GARRELS, J.I., FUTCHER, B., KOBAYASHI, R., LATTER, I., SCHWENDER, B., VOLPE, T., WARNER, J.R. AND MCLAUGHLIN, C.S. (1994). Protein identification for a *Saccharomyces cerevisiae* protein database. *Electrophoresis*, **15**, 1466–1486.
- GELFI, C., BOSSI, M.L., BJELLQVIST, B. AND RIGHETTI, P.G. (1987). Isoelectric focusing in immobilized pH gradients in the pH 10–11 range. *Journal of Biochemical and Biophysical Research Methods*, **15**, 41–48.
- GHARAHDAZHI, F., ATHERTON, D., DEMOTT, M. AND MISCHKE, S.M. (1992). Amino acid analysis of PVDF-bound proteins. in *Techniques in Protein Chemistry III* (R.H. Ageletti, Ed.), pp 249–260. Academic Press, San Diego.
- GHOSH, P., OKOH, C., LIU, Q.H. AND LAKSHMAN, M.R. (1993). Effects of chronic ethanol on enzymes regulating sialylation and desialylation of transferrin in rats. *Alcoholism: Clinical and Experimental Research*, **17**, 576–579.
- GIOMETTI, C.S., GEMMELL, M.A., TOLLAKSEN, S.L. AND TAYLOR, J. (1991). Quantitation of human leukocyte proteins after silver staining: a study with two-dimensional electrophoresis. *Electrophoresis*, **12**, 536–543.
- GIOMETTI, C.S., TAYLOR, J. AND TOLLAKSEN, S.L. (1992). Mouse liver protein database: a catalog of proteins detected by two-dimensional gel electrophoresis. *Electrophoresis*, **13**, 970–991.
- GOLAZ, O., HUGHES, G.J., FRUTIGER, S., PAQUET, N., BAIROCH, A., PASQUALI, C., SANCHEZ, J.C., TISSOT, J.D., APPEL, R.D., WALZER, C., BALANT, L. AND HOCHSTRASSER, D.F. (1993). Plasma and red blood cell protein maps: update 1993. *Electrophoresis*, **14**, 1223–1231.
- GOLD, M.R., YUNGWIRTH, T., SUTHERLAND, C.L., INGHAM, R.J., VIANZON, D., CHIU, R., VAN-OOSTVEEN, I., MORRISON, H.D. AND AEBERSOLD, R. (1994). Purification and identification of tyrosine-phosphorylated proteins from lymphocytes stimulated through the antigen receptor. *Electrophoresis*, **15**, 441–453.
- GOLDBERG, H.A., DOMENICUCCI, C., PRINGLE, G.A. AND SODEK, J. (1988). Mineral-binding proteoglycans of fetal porcine calvarial bone. *Journal of Biological Chemistry*, **263**, 12092–12101.
- GOOLEY, A.A., MARSHCHALEK, R. AND WILLIAMS, K.L. (1992). Size polymorphisms due to changes in the number of O-glycosylated tandem repeats in the *Dictyostelium discoideum* glycoprotein PsA. *Genetics*, **130**, 749–756.
- GORG, A., POSTEL, W. AND GUNTHER, S. (1988). The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **9**, 531–546.
- GORG, A., POSTEL, W., GUNTHER, S., WESER, J., STRAHLER, J.R., HANASHI, S.M., SOMERLOT, L.

- AND KUICK, R. (1988). Approach to stationary two-dimensional pattern: influence of focusing time and immobilized carrier ampholyte concentrations. *Electrophoresis*, **9**, 37-46.
- GRAVEL, P., GOLAZ, O., WALZER, C., HOCHSTRASSER, D.F., TURLER, H., AND BALANT, L.P. (1994). Analysis of glycoproteins separated by two-dimensional gel electrophoresis using lectin blotting revealed by chemiluminescence. *Analytical Biochemistry*, **221**, 66-71.
- GUNTHER, S., POSTEL, W., WIERING, H. AND GORG, A. (1988). Acid phosphatase typing for breeding nematode-resistant tomatoes by isoelectric focusing with an ultranarrow immobilized pH gradient. *Electrophoresis*, **9**, 618-620.
- HANASH, S.M., STRAHLER, J.R., NEEL, J.V., HAILAT, N., MELHEM, R., KEIM, D., ZHU, X.X., WAGNER, D., GAGE, D.A. AND WATSON, J.T. (1991). Highly resolving two-dimensional gels for protein sequencing. *Proceedings of the National Academy of Sciences USA*, **88**, 5709-5713.
- HARRINGTON, M.G., COFFMAN, J.A., CALZONE, F.J., HOOD, L.E., BRITTEN, R.J. AND DAVIDSON, E.H. (1992). Complexity of sea urchin embryo nuclear proteins that contain basic domains. *Proceedings of the National Academy of Sciences USA*, **89**, 6252-6256.
- HARRINGTON, M.G., LEE, K.H., YUN, M., ZEWEIT, T., BAILEY, J.E. AND HOOD, L.E. (1993). Mechanical precision in two-dimensional electrophoresis can improve spot positional reproducibility. *Applied and Theoretical Electrophoresis*, **3**, 347-353.
- HELLMAN, U., WERNSTEDT, C., GONEZ, J. AND HELDIN, C.-H. (1995). Improvement of an in-gel digestion for the micropreparation of internal protein fragments for amino acid sequencing. *Analytical Biochemistry*, **224**, 451-455.
- HENZEL, W.J., BILLECI, T.M., STULTS, J.T., WONG, S.C., GRIMLEY, C. AND WATANABE, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences USA*, **90**, 5011-5015.
- HESS, D., COVEY, T.C., WINZ, R., BROWNSEY, R.W. AND AEBERSOLD, R. (1993). Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Science*, **2**, 1342-1351.
- HOBOM, U., HOUTHAEVE, T. AND SANDER, C. (1994). Amino acid analysis and protein database compositional search as a rapid and inexpensive method to identify proteins. *Analytical Biochemistry*, **222**, 202-209.
- HOCHSTRASSER, D.F. AND MERRIL, C.R. (1988). 'Catalysts' for polyacrylamide gel polymerization and detection of proteins by silver staining. *Applied and Theoretical Electrophoresis*, **1**, 35-40.
- HOCHSTRASSER, D.F., PATCHORNIK, A. AND MERRIL, C.R. (1988). Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. *Analytical Biochemistry*, **173**, 412-423.
- HOCHSTRASSER, A.C., JAMES, R.W., POMETTA, D. AND HOCHSTRASSER, D.F. (1991a). Preparative isoelectrofocusing and high resolution two-dimensional electrophoresis for concentration and purification of proteins. *Applied and Theoretical Electrophoresis*, **1**, 333-337.
- HOCHSTRASSER, D.F., APPEL, R.D., VARGAS, R., PERKIER, R., VUROL, J.F., RAVIER, F., PASQUALI, C., FUNK, M., PELLIGRINI, C., MÜLLER, A.F. AND SCHERRER, J.R. (1991b). A clinical molecular scanner: the Melanie project. *Medical Computing*, **8**, 85-91.
- HOCHSTRASSER, D.F., FRUTIGER, S., PAQUET, N., BAIROCH, A., RAVIER, F., PASQUALI, C., SANCHEZ, J.-C., TISSOT, J.-D., BJELLQVIST, B., VARGAS, R., APPEL, R.D. AND HUGHES, G.J. (1992). Human liver protein map: a reference database established by microsequencing and gel comparison. *Electrophoresis*, **13**, 992-1001.
- HOLT, T.G., CHANG, C., LAURENT-WINTER, C., MURAKAMI, T., DAVIES, J.E. AND THOMPSON, C.J. (1992). Global changes in gene expression related to antibiotic synthesis in *Streptomyces hygroscopicus*. *Molecular Microbiology*, **6**, 969-980.
- HONORE, B., LEFFERS, H., MADSEN, P. AND CELIS, J.E. (1993). Interferon-gamma up-regulates a unique set of proteins in human keratinocytes. Molecular cloning and expression of the cDNA encoding the RGD-sequence containing protein IGUP I-5111. *European Journal of Biochemistry*, **218**, 421-430.
- HUBERTY, M.C., VATH, J.E., YU, W. AND MARTIN, S.A. (1993). Site-specific carbohydrate

- identification in recombinant proteins using MALD-TOF MS. *Analytical Chemistry*, **65**, 2791-2800.
- HUGHES, G.J., FRUTIGER, S., PAQUET, N., PASQUALI, C., SANCHEZ, J.-C., TISSOT, J.D., BAIROCH, A., APPEL, R.D. AND HOCHSTRASSER, D.F. (1993). Human liver protein map: update 1993. *Electrophoresis*, **14**, 1216-1222.
- HUGHES, J.H., MACK, K. AND HAMPARIAN, V.V. (1988). India ink staining of proteins on nylon and hydrophobic membranes. *Analytical Biochemistry*, **173**, 18-25.
- JAMES, P., QUADRONI, M., CARAFOLI, E. AND GONNET, G. (1993). Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, **195**, 58-64.
- JIL, H., WHITEHEAD, R.H., REID, G.E., MORITZ, R.L., WARD, L.D. AND SIMPSON, R.J. (1994). Two-dimensional electrophoretic analysis of proteins expressed by normal and cancerous human crypts: application of mass spectrometry to peptide-mass fingerprinting. *Electrophoresis*, **15**, 391-405.
- JOHNSON, R.S. AND WALSH, K.A. (1992). Sequence analysis of peptide mixtures by automated integration of Edman and mass spectrometric data. *Protein Science*, **1**, 1083-1091.
- JOHNSTON, R.F., PICKETT, S.C. AND BARKER, D.L. (1990). Autoradiography using storage phosphor technology. *Electrophoresis*, **11**, 355-360.
- JUNGBLUT, P., DZIONARA, M., KLOSE, J. AND WITTMANN-LEIBOLD, B. (1992). Identification of tissue proteins by amino acid analysis after purification by two-dimensional electrophoresis. *Journal of Protein Chemistry*, **11**, 603-612.
- JUNGBLUT, P., OTTO, A., ZEINDL-EBERHART, E., PLEIBNER, K.-P., KNECHT, M., REGITZ-ZAGROSEK, V., FLECK, E. AND WITTMANN-LEIBOLD, B. (1994). Protein composition of the human heart: the construction of a myocardial two-dimensional electrophoresis database. *Electrophoresis*, **15**, 685-707.
- KOHARA, Y., AKIYAMA, K. AND ISONO, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, **50**, 495-508.
- KLOSE, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for individual point mutations in mammals. *Human Genetik*, **26**, 231-243.
- LATHAM, K.E., GARRELS, J.I., CHANG, C. AND SOLTER, D. (1991). Quantitative analysis of protein synthesis in mouse embryos I: extensive re-programming at the one- and two-cell stages. *Development*, **2**, 921-932.
- LATHAM, K.E., GARRELS, J.I., CHANG, C. AND SOLTER, D. (1992). Analysis of embryonic mouse development: construction of a high-resolution, two-dimensional gel protein database. *Applied and Theoretical Electrophoresis*, **2**, 163-170.
- LATHAM, K.E., GARRELS, J.I. AND SOLTER, D. (1993). Two-dimensional analysis of protein synthesis. *Methods in Enzymology*, **255**, 473-489.
- LE MAIRE, M., DESCHAMPS, S., MOLLER, J.V., LE CAER, J.P. AND ROSSIER, J. (1993). Electrospray ionization mass spectrometry from sodium dodecyl sulfate-polyacrylamide gel electrophoresis: application to the topology of the sarcoplasmic reticulum  $\text{Ca}^{2+}$  ATPase. *Analytical Biochemistry*, **214**, 50-57.
- LEMKIN, P.F. AND LESTER, E.P. (1989). Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modelling. *Electrophoresis*, **10**, 122-140.
- LEMKIN, P.F., WU, Y. AND UPTON, K. (1993). An efficient disk-based data structure for rapid searching of quantitative two-dimensional gel databases. *Electrophoresis*, **14**, 1341-1350.
- LI, K.W., GERAERTS, W.P., VAN-ELK, R. AND KOOSE, J. (1989). Quantification of proteins in the subnanogram and nanogram range: comparison of the AuroDye, FerriDye, and india ink staining methods. *Analytical Biochemistry*, **182**, 44-47.
- LIANG, P. AND PARDEE, A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967-971.
- MANN, M. (1995). Sequence database searching by mass spectrometric data. In *Microcharacterisation of Proteins* (R. Kellner, F. Lottspeich, and H.E. Meyer, Eds), pp 223-245. VCH, Weinheim.

- MANN, M., HOJRUP, P. AND ROEPSTORFF, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, **22**, 338–345.
- MANN, M. AND WILKINS, M. (1994). Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**, 4390–4399.
- MATSUDAIRA, P. (1987). Sequence of picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *Journal of Biological Chemistry*, **262**, 10035–10038.
- MONARDO, P.J., BOUTELL, T., GARRELS, J.I. AND LATTER, G.I. (1994). A distributed system for two-dimensional gel analysis. *Computer Applications in the Biosciences*, **10**, 137–143.
- MORTZ, E., VORM, O., MANN, M. AND ROEPSTORFF, P. (1994). Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biological Mass Spectrometry*, **23**, 249–261.
- MURTHY, L.R. AND IQBAL, K. (1991). Measurement of picomoles of phosphoamino acids by high performance liquid chromatography. *Analytical Biochemistry*, **193**, 299–305.
- MYRICK, J.E., LEMKIN, P.F., ROBINSON, M.K. AND UPTON, K.M. (1993). Comparison of the BioImage Visage 2000 and the GELLAB-II two-dimensional electrophoresis image analysis systems. *Applied and Theoretical Electrophoresis*, **3**, 335–346.
- NEIDHARDT, F.C., APPLEBY, D.B., SANKAR, P., HUTTON, M.E. AND PHILLIPS, T.A. (1989). Genomically linked cellular protein databases derived from two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis*, **10**, 116–122.
- NIKODEM, V. AND FRESCO, J.R. (1979). Protein fingerprinting by SDS-gel electrophoresis after partial fragmentation with CNBr. *Analytical Biochemistry*, **97**, 382–386.
- NOKIHARA, K., MORITA, N. AND KURIKI, T. (1992). Applications of an automated apparatus for two-dimensional electrophoresis. Model TEP-1, for microsequence analysis of proteins. *Electrophoresis*, **13**, 701–707.
- O'FARRELL, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, **250**, 4007–4021.
- O'FARRELL, P.Z., GOODMAN, H.M. AND O'FARRELL, P.H. (1977). High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell*, **12**, 1133–1142.
- OLIVER *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46.
- OLSEN, A.D. AND MILLER, M.J. (1988). Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Analytical Biochemistry*, **169**, 49–70.
- ORTIZ, M.L., CALERO, M., FERNANDEZ-PATRON, C., PATRON, C.F., CASTELLANOS, L. AND MENDEZ, E. (1992). Imidazole-SDS-Zn reverse staining of proteins in gels containing or not SDS and microsequence of individual unmodified electroblotted proteins. *FEBS Letters*, **296**, 300–304.
- OSTERGREN, K., ERIKSSON, G. AND BJELLQVIST, B. (1988). The influence of support material used on band sharpness in Immobiline gels. *Journal of Biochemical and Biophysical Methods*, **16**, 165–170.
- OU, K., WILKINS, M.R., YAN, J.X., GOOLEY, A.A., FUNG, Y., SHEUMACK, D. AND WILLIAMS, K.L. (1995). Improved high-performance liquid chromatography of amino acids derivatised with 9-fluorenylmethyl chloroformate. *Journal of Chromatography* (in press).
- PACKER, N., WILKINS, M.R., GOLAZ, O., LAWSON, M., GOOLEY, A.A., HOCHSTRASSER, D.F., REDMOND, J. AND WILLIAMS, K.L. (1995). Characterisation of human plasma glycoproteins separated by two-dimensional gel electrophoresis. *Bio/Technology* (in press).
- PAPPIN, D.J.C., HOJRUP, P. AND BLEASBY, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, **3**, 327–332.
- PATTERSON, S.D. (1994). From electrophoretically separated protein to identification: strategies for sequence and mass analysis. *Analytical Biochemistry*, **221**, 1–15.
- PATTERSON, S.D. AND LATTER, G.I. (1993). Evaluation of storage phosphor imaging for quantitative analysis of 2-D gels using the Quest II system. *BioTechniques*, **15**, 1076–1083.
- PISANO, A., REDMOND, J.W., WILLIAMS, K.L. AND GOOLEY, A.A. (1993). Glycosylation sites identified by solid-phase Edman degradation: O-linked glycosylation motifs on human glycophorin A. *Glycobiology*, **3**, 429–435.
- RABILLOU, T. (1992). A comparison between low background silver diamine and silver nitrate protein stains. *Electrophoresis*, **13**, 429–439.

RASME  
H  
P  
RASME  
S  
A  
R  
RICH  
P  
R  
RIGH  
7  
E  
RIGH  
I  
RODR  
S  
F  
ROSE  
C  
S  
SANC  
I  
t  
SANC  
SCHE  
SHAV  
SIBB  
SIMP  
SINH  
SMIT  
STRI  
SUT  
TOU  
TOV  
URV  
VA

- RASMUSSEN, H.H., VAN DAMME, J., PUYPE, M., GESSER, B., CELIS, J.E. AND VANDEKERCKHOVE, J. (1992). Microsequences of 145 proteins recorded in the two-dimensional gel protein database of normal human epidermal keratinocytes. *Electrophoresis*, **13**, 960-969.
- RASMUSSEN, H.H., MORTZ, E., MANN, M., ROEPSTORFF, P. AND CELIS, J.E. (1994). Identification of transformation sensitive proteins recorded in human two-dimensional gel protein databases by mass-spectrometric peptide mapping alone and in combination with microsequencing. *Electrophoresis*, **15**, 406-416.
- RICHARDSON, F.C., HORN, D.M. AND ANDERSON, N.L. (1994). Dose-responses in rat hepatic protein modification and expression following exposure to the rat hepatocarcinogen methapyrilene. *Carcinogenesis*, **15**, 325-329.
- RIGHETTI, P.G. (1990). Immobilized pH gradients: theory and methodology. In *Laboratory Techniques in Biochemistry and Molecular Biology* (R.H. Burdon and P.H. van Knippenberg, Eds) Elsevier, Amsterdam.
- RIGHETTI, P.G. AND DRYSDALE, J.W. (1973). *Annals of the New York Academy of Sciences*, **209**, 163-186.
- RODRIGUEZ, L.V., GERNSTEN, D.M., RAMAGLI, L.S. AND JOHNSTON, D.A. (1993). Towards stoichiometric silver staining of proteins resolved in complex two-dimensional electrophoresis gels: real-time analysis of pattern development. *Electrophoresis*, **14**, 628-637.
- ROSENFELD, J., CAPDEVIELLE, J., GUILLEMOT, J.C. AND FERRARA, P. (1992). In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Analytical Biochemistry*, **203**, 173-179.
- SANCHEZ, J.C., RAVIER, F., PASQUALI, C., FRUTIGER, S., PAQUET, N., BJELLQVIST, B., HOCHSTRASSER, D.F. AND HUGHES, G.J. (1992). Improving the detection of proteins after transfer to polyvinylidene difluoride membranes. *Electrophoresis*, **13**, 715-717.
- SANGER, F., COULSON, A.R., HONG, G.F., HILL, D.F. AND PETERSEN, G.B. (1982). Nucleotide sequence of bacteriophage  $\lambda$  DNA. *Journal of Molecular Biology*, **162**, 729-773.
- SCHEELE, G.J. (1975). Two-dimensional analysis of soluble proteins. *Biochemistry*, **250**, 5375-5385.
- SHAW, G. (1993). Rapid identification of proteins. *Proceedings of the National Academy of Sciences USA*, **90**, 5138-5142.
- SIBBALD, P.R., SOMMERFELDT, H. AND ARGOS, P. (1991). Identification of proteins in sequence databases from amino acid composition. *Analytical Biochemistry*, **198**, 330-333.
- SIMPSON, R.J., TSUGITA, A., CELIS, J.E., GARRELS, J.I. AND MEWES, H.W. (1992). Workshop on two-dimensional gel protein databases. *Electrophoresis*, **13**, 1055-1061.
- SINHA, P.K., KOTTGEN, E., STOFFLER, M.-M., GIANAZZA, E. AND RIGHETTI, P.G. (1990). Two-dimensional maps in very acidic immobilized pH gradients. *Journal of Biochemical and Biophysical Methods*, **20**, 345-352.
- SMITH, D.W. (1994). Introduction. In *Biocomputing: Informatics and Genome Projects* (D.W. Smith, Ed.), pp1-12. Academic Press, San Diego.
- STRUPAT, K., KARAS, M., HILLENKAMP, F., ECKERSKORN, C. AND LOTTSPEICH, F. (1994). Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Analytical Chemistry*, **66**, 464-470.
- SUTTON, C.W., PEMBERTON, K.S., COTTRELL, J.S., CORBETT, J.M., WHEELER, C.H., DUNN, M.J. AND PAPPIN, D.J. (1995). Identification of myocardial proteins from two-dimensional gels by peptide mass fingerprinting. *Electrophoresis*, **16**, 308-316.
- TOUS, G.I., FAUSNAUGH, J.L., AKINYOSOYE, O., LACKLAND, H., WINTERCASH, P., VITORICA, F.J. AND STEIN, S. (1989). Amino acid analysis on polyvinylidene difluoride membranes. *Analytical Biochemistry*, **179**, 50-55.
- TOVEY, E.R., FORD, S.A. AND BALDO, B.A. (1987). Protein blotting on nitrocellulose: some important aspects of the resolution and detection of antigens in complex extracts. *Journal of Biochemical and Biophysical Methods*, **14**, 1-17.
- URWIN, V.E. AND JACKSON, P. (1993). Two-dimensional polyacrylamide gel electrophoresis of proteins labeled with the fluorophore monobromobimane prior to first-dimensional isoelectric focusing: imaging of the fluorescent protein spot patterns using a cooled charge-coupled device. *Analytical Biochemistry*, **209**, 57-62.
- VANBOGELEN, R.A., HUTTON, M.E. AND NEIDHARDT, F.C. (1990). Gene-protein database

- of *Escherichia coli* K-12: edition 3. *Electrophoresis*, **11**, 1131–1166.
- VANBOGELEN, R.A. AND NEIDHARDT, F.C. (1991). The gene-protein database of *Escherichia coli*: edition 4. *Electrophoresis*, **12**, 955–994.
- VANBOGELEN, R.A., SANKER, P., CLARK, R.L., BOGAN, J.A. AND NEIDHARDT, F.C. (1992). The gene-protein database of *Escherichia coli*: edition 5. *Electrophoresis*, **13**, 1014–1054.
- VANDEKERKHOVE, J., BAUW, G., VANCOMPERNOLLE, K., HONORE, B. AND CELIS, J. (1990). Comparative two-dimensional gel analysis and microsequencing identifies gelsolin as one of the most prominent downregulated markers of transformed human fibroblast and epithelial cells. *Journal of Cell Biology*, **111**, 95–102.
- VANFLETEREN, J.R., RAYMACKERS, J.G., VAN BUN, S.M. AND MEHUS, L.A. (1992). Peptide mapping and microsequencing of proteins separated by SDS-PAGE after limited *in situ* hydrolysis. *BioTechniques*, **12**, 550–557.
- VORM, O. AND MANN, M. (1994). Improved mass accuracy in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of peptides. *Journal of the American Society for Mass Spectrometry*, **5**, 955–958.
- VORM, O., ROEPSTORFF, P. AND MANN, M. (1994). Improved resolution and very high sensitivity in MALDI TOF of matrix surfaces made by fast evaporation. *Analytical Chemistry*, **66**, 3281–3287.
- WALLACE, A. AND SALUZ, H.P. (1992a). Ultramicrodetection of proteins in polyacrylamide gels. *Analytical Biochemistry*, **203**, 27–34.
- WALLACE, A. AND SALUZ, H.P. (1992b). Beyond silver staining. *Nature*, **357**, 608–609.
- WALSH, B.J., GOOLEY, A.A., WILLIAMS, K.L. AND BREIT, S.N. (1995). Identification of macrophage activation associated proteins by two-dimensional electrophoresis and microsequencing. *Journal of Leukocyte Biology*, **57**, 507–512.
- WASINGER, V.C., CORDWELL, S.J., POLJAK, A., YAN, J.X., GOOLEY, A.A., WILKINS, M.R., DUNCAN, M., HARRIS, R., WILLIAMS, K.L. AND HUMPHERY-SMITH, I. (1995). Progress with Gene-Product Mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*, **16**. In Press.
- WEITZHANDLER, M., KADLECEK, D., AVDALOVIC, N., FORTE, J.G., CHOW, D. AND TOWNSEND, R. R. (1993). Monosaccharide and oligosaccharide analysis of proteins transferred to polyvinylidene fluoride membranes after sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry*, **268**, 5121–5130.
- WILKINS, M.R., PASQUALI, C., APPEL, R.D., OU, K., GOLAZ, O., SANCHEZ, J-C., YAN, J.X., GOOLEY, A.A., HUGHES, G., HUMPHERY-SMITH, I., WILLIAMS, K.L. AND HOCHSTRASSER, D.F. (1995). From Proteins to Proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Submitted.
- WILKINS, M.R., OU, K., APPEL, R.D., GOLAZ, O., PASQUALI, C., YAN, J.X., FARNSWORTH, V., CARTIER, P., HOCHSTRASSER, D.F., WILLIAMS, K.L. AND GOOLEY, A.A. (1996). Rapid protein identification using N-terminal sequence tagging and amino acid analysis (submitted).
- WIRTH, P.J., LUO, L.D., FUJIMOTO, Y., BISGAARD, H.C. AND OLSEN A.D. (1991). The rat liver epithelial (RLE) cell protein database. *Electrophoresis*, **12**, 931–954.
- WIRTH, P.J., LUO, L.D., BENJAMIN, T., HOANG, T.N., OLSEN A.D. AND PARMALEE, D.C. (1993). The rat liver epithelial (RLE) cell nuclear protein database. *Electrophoresis*, **14**, 1199–1215.
- WU, Y., LEMKIN, P.F. AND UPTON, K. (1993). A fast spot segmentation algorithm for two-dimensional gel electrophoresis analysis. *Electrophoresis*, **14**, 1351–1356.
- YAMAGUCHI, K. AND ASAKAWA, H. (1988). Preparation of colloidal gold for staining proteins electrotransferred onto nitrocellulose membranes. *Analytical Biochemistry*, **172**, 104–107.
- YAMASHITA, K., IDEO, H., OHKURA, T., FUKUSHIMA, K., YUASA, I., OHNO, K. AND TAKESHITA, K. (1993). Sugar chains of serum transferrin from patients with carbohydrate deficient glycoprotein syndrome. Evidence of asparagine-N-linked oligosaccharide transfer deficiency. *Journal of Biological Chemistry*, **268**, 5783–5789.
- YATES, J.R. III, SPEICHER, S., GRIFFIN, P.R. AND HUNKAPILLER, T. (1993). Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry*, **214**, 397–408.

### 3 The

R.M. D

<sup>1</sup> *Therm*  
*Hamilt*  
*Wales 2*  
*Private*

### Introdu

Proteas  
dealing  
type. T  
structur  
since th  
by Was  
1930) c  
possibly  
mechan  
subtilis  
1986; f  
comme  
proteas  
the cha  
been in  
regulat  
apparer  
Kalisz.  
Barrett.  
Give  
surpris  
attracti  
subject  
stabilit  
enzyme  
Sherod

\* Corres.  
*Biotechn*  
0264–872

## Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing

JULIO E. CELIS,\*<sup>1</sup> HANNE H. RASMUSSEN,\* HENRIK LEFFERS,\* PEDER MADSEN,\* BENT HONORÉ,\* BORBALA GESSER,\* KURT DEJGAARD,\* JOËL VANDEKERCKHOVE<sup>†</sup>

<sup>\*</sup>Institute of Medical Biochemistry and Human Genome Research Centre, Aarhus University, DK-8000 Aarhus, Denmark and <sup>†</sup>Laboratorium voor Fysiologische Chemie, Rijksuniversiteit Gent, Belgium

**ABSTRACT** Analysis of cellular protein patterns by computer-aided 2-dimensional gel electrophoresis together with recent advances in protein sequence analysis have made possible the establishment of comprehensive 2-dimensional gel protein databases that may link protein and DNA information and that offer a global approach to the study of the cell. Using the integrated approach offered by 2-dimensional gel protein databases it is now possible to reveal phenotype specific protein (or proteins), to microsequence them, to search for homology with previously identified proteins, to clone the cDNAs, to assign partial protein sequence to genes for which the full DNA sequence and the chromosome location is known, and to study the regulatory properties and function of groups of proteins that are coordinately expressed in a given biological process. Human 2-dimensional gel protein databases are becoming increasingly important in view of the concerted effort to map and sequence the entire genome. — Celis, J. E.; Rasmussen, H. H.; Leffers, H.; Madsen, P.; Honoré, B.; Gesser, B.; Dejgaard, K.; Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing. *FASEB J.* 5: 2200-2208; 1991.

**Key Words:** human protein patterns • 2-dimensional gel protein databases • gene expression • microsequencing • cDNA cloning • linking protein and DNA information • genome mapping and sequencing

PROTEINS SYNTHESIZED FROM information contained in the DNA orchestrate most cellular functions. The total number of proteins synthesized by a typical human cell is unknown although current estimates range from 3000 to 6000. Of these, as many as 70% may perform household functions and are expected to be shared by all cell types irrespective of their origin. There are many different cell types in the human body with perhaps 30,000 to 50,000 proteins expressed in the organism as a whole judged from the fact that about 3% of the haploid genome correspond to genes. Today only a small fraction of the total set of proteins has been identified, and little is known about the protein patterns of individual cell types or their variation under physiological and abnormal conditions.

For the past 15 years, high resolution 2-dimensional gel electrophoresis has been the technique of choice to determine the protein composition of a given cell type and for monitoring changes in gene activity through quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions (refs 1-6 and references

therein). The technique originally described by O'Farrell (1) separates proteins in terms of their isoelectric point (pI) and molecular weight. Usually one chooses a condition of interest and the cell reveals the global protein behavioral response as all detected proteins can be analyzed both qualitatively and quantitatively in relation to each other. At present, most available 2-dimensional gel techniques (regular gel format) can resolve between 1000 and 2000 proteins from a given mammalian cell type, a number that corresponds to about 2 million base pairs of coded DNA. Less abundant proteins can be detected by analyzing partially purified cellular fractions.

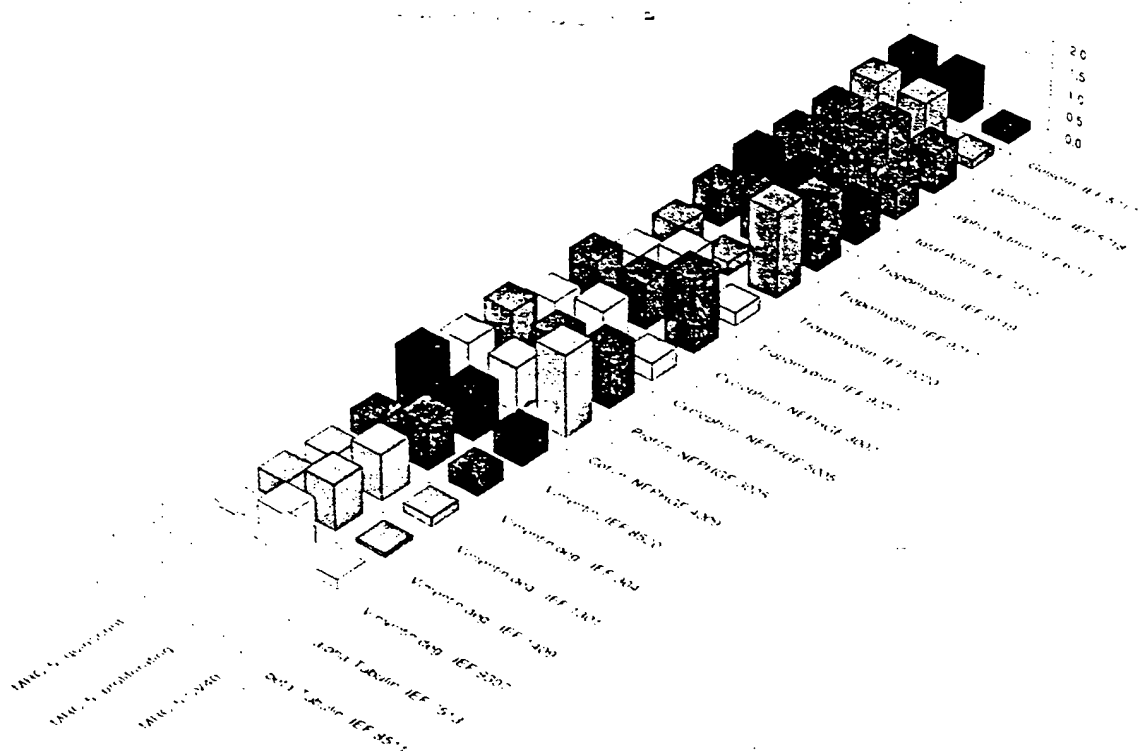
Two-dimensional gel electrophoresis has been widely applied to analysis of cellular protein patterns from bacteria to mammalian cells (refs 1-6, and references therein). In spite of much work, however, information gathered from these studies has not reached the scientific community in its fullness because of lack of standardized gel systems and the lack of means for storing and communicating protein information. Only recently, because of the development of appropriate computer software (7-13), has it been possible to scan gels, assign numbers to individual proteins, and store the wealth of information in quantitative and qualitative comprehensive 2-dimensional gel protein databases (4, 14-23), i.e., those containing information about the various properties (physical, chemical, biological, biochemical, physiological, genetic, immunological, architectural, etc.) of all the proteins that can be detected in a given cell type. Such integrated 2-dimensional gel protein databases offer an easy and standardized medium in which to store and communicate protein information and provide a unique framework in which to focus a multidisciplinary approach to study the cell. Once a protein is identified in the database, all of the information accumulated can be easily retrieved and made available to the researcher. In the long run, protein databases are expected to foster a wide variety of biological information that may be instrumental to researchers working in many areas of biology—among others, cancer and oncogene studies, differentiation, development, drug development and testing, genetic variation, and diagnosis of genetic and clinical diseases (Fig. 1).

The approach using systematic 2-dimensional gel protein analysis has recently gained a new dimension with the advent of techniques to microsequence major proteins recorded

<sup>1</sup>To whom correspondence should be addressed, at: Institute of Medical Biochemistry and Human Genome Research Centre, Ole Worms Alle, Bldg. 170, University Park, DK-8000 Aarhus C, Denmark.

# THE FASEB JOURNAL

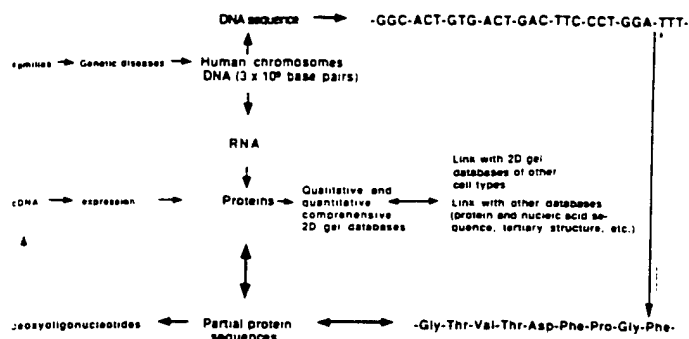
Official Publication of the Federation of American Societies for Experimental Biology  
APS ASBMB ASPET AAP AIN AAI ASCB



## 2-D GEL ELECTROPHORESIS AND MICROSEQUENCING

In this issue: Non-insulin-dependent diabetes • Trafficking of insulin receptor  
n-3 Polyunsaturated fatty acid effects • Matrix metalloproteinases  
G-protein-coupled ion channel pathways • RNA minihelices





**Figure 1.** Interface between partial protein sequence databases, comprehensive 2-dimensional gel databases, and the human genome sequencing project. Appropriate software is required to compare protein and DNA sequences. In general, although the inference of a protein's sequence from the DNA sequence (thick arrow) is direct and unambiguous, the DNA sequence can only be inferred approximately from the protein sequence (thin arrow) and cloning of the gene requires either a cDNA or the requisite group of oligonucleotide probes deduced from the partial amino acid sequence. Modified from ref 6.

in the databases (refs 24-42 and references therein). Partial protein sequences can be used to search for protein identity as well as to prepare specific DNA probes for cloning as yet-uncharacterized proteins (Fig. 1). As these sequences can be stored in the database (see for example Fig. 2H), they offer a unique opportunity to link information on proteins with the existing or forthcoming DNA sequence data on the human genome (Fig. 1) (20, 36, 39).

Using the integrated approach offered by comprehensive 2-dimensional gel databases (Fig. 1), it will be possible to identify phenotype-specific proteins; microsequence them and store the information in the database; search for homology with previously characterized proteins; clone the cDNAs, assign partial protein sequences to genes for which the full DNA sequence and the chromosome location are known, and study the regulatory properties and function of groups of proteins (pathways, organelles, etc.) that are coordinately expressed in a given biological process. Comprehensive 2-dimensional gel protein databases will depict an integrated picture of the expression levels and properties of the thousands of protein components of organelles, pathways, and cytoskeletal systems in both physiological and abnormal conditions and are expected to lead to identification of new regulatory networks in different cell types and organisms. In the future, 2-dimensional gel protein databases may be linked to each other as well as to national and international specialized databanks on nucleic acid and protein sequences, protein structures, NMR experimental data, complex carbohydrates, etc.

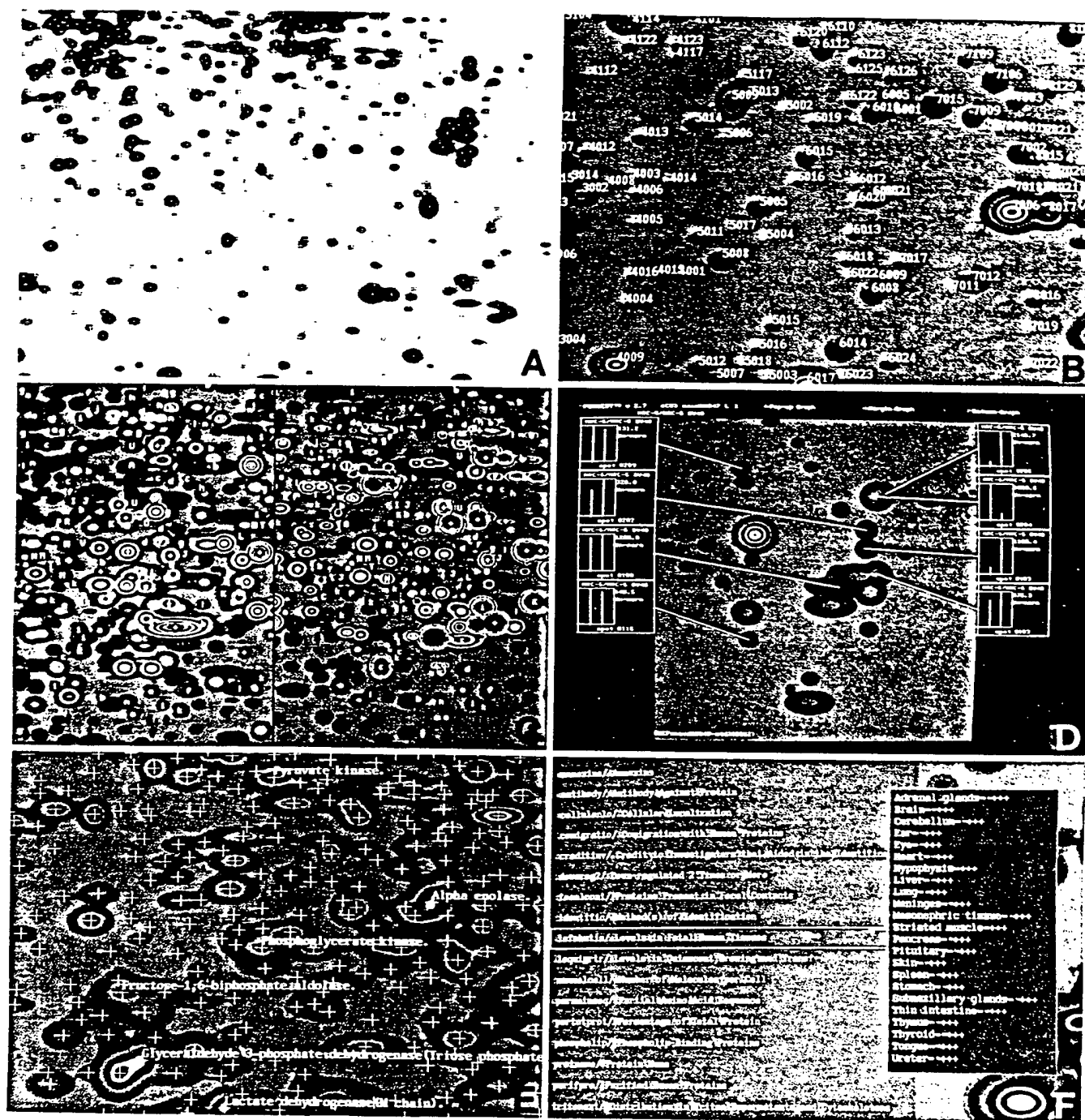
A few 2-dimensional gel protein databases that are accessible in a computer form have been published in extenso: these correspond to the protein-gene database of *Escherichia coli* K-12 developed by Neidhardt and colleagues (14, 23), the rat REF 52 database established by Garrels and co-workers at Cold Spring Harbor (18, 22), and a few human databases (transformed amnion cells [15, 20], normal embryonal lung MRC-5 fibroblasts [17, 21], keratinocytes [19] and peripheral blood mononuclear cells [15]) developed in Aarhus. Given space limitations and to keep this review in focus, we will concentrate on the computerized analysis of human cellular 2-dimensional gel patterns, and in particular on the steps involved in establishing comprehensive 2-dimensional gel databases that can link protein and DNA information.

## MAKING AND MANAGING A COMPREHENSIVE 2-DIMENSIONAL GEL DATABASE OF HUMAN CELLULAR PROTEINS

The first step in making a comprehensive 2-dimensional gel protein database is to prepare a synthetic image (digital form of the gel image) of the gel (fluorogram, Coomassie blue or silver stained gel) to be used as a standard or master reference. This can be done with laser scanners, charge couple device (CCD)<sup>2</sup> array scanners, television cameras, rotating drum scanners, and multiwire chambers (13). Computerized analysis systems for spot detection, quantitation, pattern matching, and data handling (access and retrieval of information, database making) have been described in the literature (ELSIE [43], GELLAB [11], HERMES [44], MELANIE [10], QUEST (9), and TYCHO [8]) and some are available commercially (PDQUEST, Protein Database Inc., Huntington, NY; KEPLER, Large Scale Biology, Rockville, Md.; Visage, BioImage Corporation, Ann Arbor, Mich.; Gemini, Joyce Loeb, Gateshead; Microscan 1000, Technology Resources Inc., Nashville, Tenn. and MasterScan, Billerica, Mass.). Unfortunately, most of these systems are incompatible with one another and their advantages and disadvantages have been discussed by Miller (13).

In our work station in Aarhus, fluorograms are scanned with a Molecular Dynamics laser scanner and the data are analyzed using the PDQUEST II software (Protein Databases Inc.) (12) running on a spark station computer 4100 FC-8-P3 from SUN Microsystems, Inc. The scanner measures intensity in the range of 0-2.0 absorbance. A typical scan of a 17 x 17 cm fluorogram takes about 2 min. Steps in image analysis include: initial smoothing, background subtraction, final smoothing, spot detection, and fitting of ideal Gaussian distribution to spot centers. Spot intensity is calculated as the integration of a fitted Gaussian. If calibration strips containing individual segments of a known amount of radioactivity are used, it is possible to merge multiple exposures of the sample image into a single data image of greater dynamic range. Once the synthetic image is created it can be stored on disk and displayed directly on the monitor. Functions that can be used to edit the images include: cancel (for example, to erase scratches that may have been interpreted as spots by the computer; cancel streaks or low dpm spots), combine (sometimes a spot may be resolved into several closely packed spots), restore, uncombine, and add spot to the gel. The process is time consuming—about 1-1/2 day per image. Edited standard images can be matched to other synthetic images. **Figure 2A** shows a portion of a standard synthetic image (IEF) of a fluorogram of [<sup>35</sup>S]methionine labeled cellular proteins from human AMA cells (master database) (20). Images can be displayed either in black and white (resembling the original fluorograms) or in color (other images in Fig. 2), depending on the need. As shown in Fig. 2B, each polypeptide is assigned a number by the computer, which facilitates the entry and retrieval of qualitative and quantitative information for any given spot in the gel (20). The standard image can be matched automatically by the computer to other standard or reference gels (Fig. 2C, matching of AMA cellular proteins [left] to MRC-5 proteins [right]) provided a few landmark spots are given manually as reference (indicated with a + in Fig. 2C) to initiate the process.

<sup>2</sup>Abbreviations: CCD, charge couple device; PCNA, proliferating cell nuclear antigen; HPLC, high performance liquid chromatography.



**Figure 2.** A) Synthetic image of a fraction of an IEF gel of the master image of AMA cellular proteins. B) As in A but showing numbers assigned to each spot. C) Comparison of AMA (left) and normal human embryonal lung MRC-5 fibroblasts (right) IEF proteins patterns. Matched proteins are indicated by a + or by the same letters in both gels. Once a protein is matched, information contained in the various categories available in the master AMA database can be transferred. D) Synthetic image of a fraction of an IEF fluorogram of [<sup>35</sup>S]methionine labeled proteins from normal human MRC-5 fibroblasts. The histograms show levels of synthesis of a few proteins in MRC-5 (left bar) and SV40 transformed MRC-5 (right bar) fibroblasts. E) Polypeptides that contain information under the category glycolytic pathway. F) The function peruse annotation for spot allows the operator to inquire about categories and information available for a given protein. G) Relative abundance of cytoskeletal and cytoskeletal-related proteins in quiescent, proliferating, and SV40-transformed MRC-5 fibroblasts. H) Polypeptides that contain information under the category partial amino acid sequences.



cross-matched experiments (18, 22).

Once a standard map of a given protein sample is made, one can enter qualitative annotations to make a reference database. Our master 2-dimensional gel database of transformed human amnion cell (AMA) proteins (20) lists 3430 polypeptides of which 2592 correspond to cellular components, having pI's ranging from 4 to 13 and molecular weights between 8.5 and 230 kDa. The most abundant proteins in the database correspond to total actin (3.87% of total protein; about 90 million molecules per cell) while the lesser abundant of the recorded polypeptides are present in the vicinity of 5000 molecules per cell. Some annotation categories we are using to establish the master AMA database include: 1) protein identification (comigration with purified proteins, 2-dimensional immunoblotting, microsequencing); 2) amounts (total amounts and levels of synthesis); 3) subcellular localization (nuclear, cytoskeletal, membrane, membrane receptors, specific organelles, etc.); 4) antibodies; 5) posttranslational modifications (phosphorylation, glycosylation, methylation etc.); 6) microsequencing; 7) cell cycle specificity (specific variations in levels of synthesis and amount); 8) regulatory behavior (effect of hormones, growth factors, heat shock, etc.); 9) rate of synthesis in normal and transformed cells (proliferation sensitive proteins, cell cycle specific proteins, oncogenes, components of the pathway (or pathways) that control cell proliferation); 10) function (mainly from comigration with proteins of known function); 11) sets of proteins that are coordinately regulated (hierarchy of controls, differential gene expression in various cells, etc.); 12) cDNAs (cloned cDNAs); 13) proteins that are specific to a given disease (systematic comparison of protein patterns of fibroblast proteins from healthy and diseased individuals); 14) expression and exploitation of transfected cDNAs; 15) pathways (metabolic, others); 16) gene localization (genetic and physical); 17) effect of microinjected antibody on patterns of protein synthesis; and 18) secreted proteins.

Information entered for any spot in a given annotation category can be easily retrieved by asking the computer to display the information on the color screen. For example, Fig. 2E shows a synthetic image of a NEPHGE gel (master AMA database) displaying the information contained under the entry glycolytic pathway. Alternatively, one can use the function peruse annotations for spot to directly ask the computer to list all the entries available for a particular protein. By clicking the mouse in a given entry (in this case, presence in fetal human tissues) it is possible to take a quick look at the information in that particular entry (Fig. 2F).

A major obstacle encountered in building comprehensive 2-dimensional gel protein databases is identifying the large number of proteins separated by this technology. In our databases (20, 21), known proteins are identified by one or a combination of the following procedures: 1) comigration with known proteins, 2) 2-dimensional gel immunoblotting using specific antibodies, and 3) microsequencing of Coomassie Brilliant Blue stained human proteins recovered from dried 2-dimensional gels (see next section). Protein identification by means of microsequencing may be difficult, as individual protein members of families with short peptide differences may escape detection. In the gene-protein database of *E. coli* K-12 (14, 23), another major 2-dimensional gel database available at present, proteins are being identified by a wider range of tests that include comigration with purified proteins; genetic criterion (deletion, insertion, frameshift, nonsense, missense, regulatory), plasmid-bearing strains and in vitro synthesis of protein; selective labeling (methylation, phosphorylation); peptide map similarity; and physiological criterion and selective derivatization.

So far we have received nearly 550 antibodies from laboratories all over the world and these are being systematically tested by 2-dimensional gel immunoblotting for antigen determination. Similarly, purified proteins and organelles provided by several laboratories have greatly aided identification of unknown proteins (20, 21). We routinely request antibodies and protein samples and promise the donors to make available all the information we may have accumulated on that particular protein. For example, Table 1 lists entries available for Lipocortin V (IEF SSP 8216), also known as annexin V, VAC- $\alpha$ , endonexin II, renocortin, chromobindin-5', anticoagulant protein, PAP-I,  $\gamma$ -calmedin, IBC, calphobindin, and anchorin CII.

As mentioned previously, one distinct advantage of 2-dimensional gel electrophoresis is the possibility of studying quantitative variations in cellular protein patterns that may lead to identification of groups of proteins that are expressed coordinately during a given biological process. Quantitation, however, is not an easy task as reflected by the lack of published data on global cellular protein patterns. We believe this is partly due to difficulties in obtaining sets of gels that are suitable for computer analysis (streaking, material remaining at the origin, etc.) as well as to limitations (laborious editing time, need of calibration strips to merge images, limited dynamic range, etc.) in the computer analysis systems available at the moment. Perhaps the most advanced quantitative studies published so far using computer analysis have been carried out by Garrels and co-workers (18, 22). In particular, these investigators have established a quantitative rat protein database (18, 22) designed to study growth control (proliferation, growth inhibitors, and stimulation) and transformation in well-defined groups of cell lines obtained by transformation of rat REF52 cells with SV40, adenovirus, and the Kirsten murine sarcoma virus. These studies have revealed clusters of proteins induced or repressed during growth to confluence as well as groups of transformation-sensitive proteins that respond in a differential fashion to transformation by DNA and RNA viruses. A most interesting feature of this quantitative database is the discovery of a group of coregulated proteins that show similar expression patterns as the cell cycle-regulated DNA replication protein known as proliferating cell nuclear antigen (PCNA)/cyclin (45).

In our human databases, most quantitations have been carried out by estimating the radioactivity contained in the polypeptides by direct counting of the gel pieces in a scintillation counter (20, 21). Up to 700 proteins can be cut out through appropriate exposed films in a period of time comparable to that required for editing a synthetic image. Manual quantitation of this large number of spots is difficult without the assistance of a master reference image and a numbering system that can be used to identify the spots. Using this approach, we have recorded quantitative changes in the relative abundance of 592 [ $^{35}$ S]methionine-labeled proteins synthesized by quiescent, proliferating, and SV40 transformed human embryonic lung MRC-5 fibroblasts (21). Some data concerning cytoskeletal and cytoskeletal-related proteins are presented in Fig. 2G. Our studies as well as those of Garrels and co-workers (18, 22) may in the long run help define patterns of gene expression that are characteristic of the transformed state.

## OTHER 2-DIMENSIONAL GEL PROTEIN DATABASES

As mentioned previously there are other 2-dimensional gel databases available in computer form that have been pub-

TABLE 1. Some entries for lipocortin V in the human A.M.A. 2-dimensional gel protein database

Entries for lipocortin V (IEF SSP 8216)	Information entered
1. Protein name	Lipocortin V, renocortin, chromobindin-5', endonexin I, anticoagulant protein, PAP-I, VAC- $\alpha$ , 35- $\gamma$ -calcimedlin, IBC, calphobindin I, anchorin CII, annexin V
2. Percentage of total protein	0.110% (about 2,800,000 molecules per cell)
3. Apparent molecular weight (mr)	33.3 kDa
4. Isoelectric point (pI)	4.76
5. Method (or methods) of identification	Microsequencing, 2-dimensional immunoblotting, Comigration
6. Credit to investigators that aided in identification	G. Bauw, J. Vandekerckhove, and colleagues, Rijksuniversiteit Gent; B. Pepinsky, BIOGEN, Cambridge; N.G. Ahn, University of Washington
7. Antibody against protein	Polyclonal (rabbit, antibody no. 20), B. Pepinsky, BIOGEN, Cambridge
8. Comigration with human proteins	Lipocortin V, N.G. Ahn, Howard Hughes Medical Institute, Washington University
9. Cellular localization	Subcortical membrane
10. Calcium/phospholipid-dependent membrane proteins	Lipocortin V
11. Function	Regulation of various aspects of inflammation, immune response, blood coagulation and differentiation
12. Partial amino acid sequence	GTVTDFPGFDER (7-18), VLTEIIASR (109-117), QVYEEYGGSSLEDDVVG (127-143), ?GTDEEKFITIFGT(R) (187-201)
13. cDNA sequence	Known, R. Blake et al., <i>J. Biol. Chem.</i> <b>263</b> , 10799-10811; 1988 (pI = 4.76 from translated sequence)
14. Levels in fetal human tissues	Adrenal glands = + + +; brain = + + +; cerebellum = + + +; ear = + + +; eye = + + +; heart = + + +; hypophysis = + + +; liver = + + +; lung = + + +; meninges = + + +; mesonephric tissue = + + +; striated muscle = + + +; pancreas = + + +; skin = + + +; spleen = + + +; stomach = + + +; submandibular gland = + + +; small intestine = + + +; thymus = + + +; thyroid gland = + + +; tongue = + + +; ureter = + + +
15. Levels in quiescent, proliferating, and transformed MRC-5 fibroblasts	Q (quiescent) = 1.1; P (proliferating) = 1.0; T (SV40 transformed) = 0.3
16. Distribution in Triton supernatant and cytoskeletons	Mainly supernatant

lished in extenso: these correspond to the *E. coli* K-12 protein-gene database (14, 23) and to the rat REF52 database (18, 22).

The *E. coli* K-12 cellular protein-gene database is perhaps the most complete of all databases reported so far and eventually it should trace each protein back to its structural gene. Information contained in this database includes: gene/protein name (protein name, EC number, gene name); 2-dimensional gel spot designations (x-y coordinates from reference gels, alphanumeric designation); genetic information (linkage map location, physical map location, Genbank code, sequence reference, location on Kohara clones); biochemical information (molecular weight, pI, number of residues of each amino acid, mole percent of each amino acid, total number of amino acids in a polypeptide), and regulatory information (cellular level of protein in different media and different temperature, member of regulon, member of stimulon). Major advances of this database are envisaged in the future in view of the eminent sequencing of

the whole *E. coli* genome as well as the development of improved methods to express cloned genes.

The rat REF52 2-dimensional gel protein database lists about 1600 proteins that have been recorded using the QUEST analysis system (18, 22). Included in this quantitative database are 1) protein names (cytoskeletal and heat shock proteins as well as various nuclear, mitochondrial, and cytoplasmic proteins), 2) annotations (subcellular localization, modification, recognition by specific antibodies, coprecipitation, NH<sub>2</sub>-terminal sequence, cross-reference to protein sequence information and references to the literature), 3) protein sets (cytoskeletal proteins, phosphoproteins, sets of proteins with PCNA/cyclin-like properties, etc.) and 4) general quantitative data (protein synthesis during growth of normal REF52 cells to confluence and quiescence, and after restimulation of growth-inhibited cells).

In addition to the 2-dimensional gel databases mentioned so far there are several smaller cellular databases being established in human (normal human diploid fibroblasts, lym-

phocytes, leukocytes, leukemic cells) mouse (NIH/3T3 cells, T lymphocytes), *Aplysia*, yeast (*Saccharomyces cerevisiae*), plants (wheat, barley, sorghum), and *Euglena*. Databases of tissue protein, (brain, whole mouse, liver) and body fluid proteins (plasma proteins, cerebrospinal fluid, urine, and milk) are being established in several laboratories. The reader is directed to the review by Celis et al. (4) for details and references concerning these databases.

#### MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE 2-DIMENSIONAL GEL DATABASES: A DIRECT LINK BETWEEN PROTEINS AND GENES

The development of highly sensitive amino acid gas-phase or liquid-phase sequencers (24), together with the establishment of efficient protein and peptide sample preparation methods, has opened the possibility to perform a systematic sequence analysis of proteins resolved by 2-dimensional gel electrophoresis. Indeed, generated pieces of protein sequences can be used to search for protein identity (comparison with available sequences stored in databanks) as well as for preparing specific DNA probes for cloning of as yet uncharacterized proteins (Fig. 1). In addition, partial protein sequences can be stored in 2-dimensional gel databases (for example, see Fig. 2H) and offer a unique link between proteins and genes (Fig. 1).

In the early 1970s gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in ref 25). Proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation at the nanomole level. This technique was further refined by using electro-elution to recover proteins and by miniaturizing the system (26). This method has been used extensively, but showed increasing drawbacks (low yields, protein samples contaminated by free amino acids, and  $\text{NH}_2$ -terminal blocking) as the amounts of handled protein gradually became smaller (e.g., at the 10 picomol level).

Most of the problems referred to above have been minimized with the introduction of protein-electroblotting procedures (27-32). When proteins are blotted on chemically inert membranes, it is possible to sequence the immobilized proteins directly without additional manipulations. Thus, depending on the amount of bound protein and its nature, this direct sequencing procedure generally yields  $\text{NH}_2$ -terminal sequences containing 10-40 residues. As such, this technique was used to identify, by their  $\text{NH}_2$ -terminal sequences, differentially expressed major proteins from total cellular extracts separated on 2-dimensional gels. A major difficulty encountered in this procedure is the occurrence of frequent artefactual blockage of the proteins. Several studies suggest that this phenomenon is mainly due to reaction with contaminants (particularly unpolymerized acrylamide present in the gel) and to a high dilution of the protein (low concentration of the protein per unit membrane surface). In addition to this primarily technical problem, many proteins are blocked in vivo by acylation or by a pyrrolidone carboxylic acid cap.

The problem of partial or complete  $\text{NH}_2$ -terminal blockage can be circumvented by generating internal amino acid sequences. This is achieved by fragmenting the protein present in the gel (gel in situ cleavage) or by cleaving it while bound to the membrane (membrane in situ cleavage) (33-35). In both cases, proteins are either cleaved in a restricted way (e.g., by limited enzymatic digestion or by using restriction chemical cleavage conditions) or fragmented into smaller peptides.

Of the different combinations examined, we had good results by using exhaustive proteolytic digestion on membrane-immobilized proteins. This method has been described for Ponceau red-stained proteins on nitrocellulose blots (34), for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass fiber membranes (35). The proteases used (trypsin, chymotrypsin, or pepsin) cleave at multiple sites, generating small peptides that elute from the blot into the digestion buffer from which they are purified by reversed-phase high performance liquid chromatography (HPLC) before being sequenced individually. Although each of these manipulations could be expected to result in a reduced yield of final sequence information, we were surprised that the peptides could be sequenced with high efficiency. In our hands, this approach could be routinely applied to gel-purified proteins available in amounts ranging from 5 to 10  $\mu\text{g}$ , and often yielded sequence information covering more than 30% of the total protein. As membrane-immobilized proteins are not homogeneously digested, but rather show protease sensitivity next to resistant regions, the number of peptides generated is much lower than expected from the number of potential cleavage sites. Consequently, HPLC peptide chromatograms are less complex and most peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a 2-dimensional gel, proteins of interest are often obtained in yields insufficient for the currently available sequencing technology. More material can be obtained by enriching for a certain subcellular fraction (purified cell organelles) or by exploiting affinity (dyes, metals, drugs, etc) or hydrophobic properties of proteins before gel analysis. All of the sequencing results accumulated so far in the human protein database (20) (a few are shown in Fig. 2H) have been obtained from analysis of protein spots collected from 2-dimensional gels that had been stained with Coomassie blue according to standard procedures and dried for storage. Proteins are recovered from the collected gel pieces by a protein-elution-concentration device, combined with gel electrophoresis and electroblotting. Details of this technique have been reported in a previous communication (42) and a brief outline is given below.

Combined gel pieces are allowed to swell in gel sample buffer (a total volume of 1.5 ml). The gel pieces combined with the supernatant are then collected into a large slot made in a new gel. The slot is further filled with Sephadex G-10 equilibrated in gel sample buffer. During consecutive gel electrophoresis, most of the electrical current passes on the side of the slot instead of passing through the slot. This results in both a vertical stacking and horizontal contraction of the protein band. With this device the protein is efficiently eluted from the gel pieces and concentrated from a large volume into a narrow spot. The highly concentrated (about 5  $\text{mm}^2$ ) protein spot is then electroblotted on PVDF-membranes, stained with Amido black, and in situ digested with trypsin. The peptides generated during digestion elute from the membrane into the supernatant, and can be separated by narrow bore reversed-phase HPLC and collected individually for sequence analysis.

Using this and previous procedures (37, 39, 42), we have so far analyzed 70 protein spots collected from 2-dimensional gels (20, and unpublished observations) (see for example Fig. 2H). The sequence information amounts to 2100 allocated residues corresponding to an average of 30 residues per protein spot. So far we have made cDNAs of many of the unknown proteins that have been microsequenced, and a substantial number has been cloned and sequenced. All available information indicates that it may be possible to obtain partial sequence information from most of

the proteins that can be visualized by Coomassie Brilliant Blue staining.

Partial protein sequences are stored in the database as displayed in Fig. 2H, and it should be possible in the near future to interface this information with forthcoming DNA sequence data from the human genome project. In the long run, as the human genome sequences become available it will be possible to assign partial protein sequences to genes for which the full DNA sequence and chromosomal location are known (Fig. 1).

## SUMMARY

The studies presented in this brief review are intended to demonstrate the usefulness of computer-aided 2-dimensional gel electrophoresis and microsequencing to analyze cellular protein patterns, and to link protein and DNA information. As more information is gathered worldwide, comprehensive databases will depict an integrated picture of the expression levels and properties of the thousands of proteins that orchestrate most cellular functions.

Clearly, databases allow easy access to a large body of data and provide an efficient medium to communicate standardized protein information. In the future, databases will foster a wide variety of biological information that can be used to support collaborative research projects in basic and applied biology as well as in clinical research (2, 5, 46). Once a protein is identified in a particular database all the information gathered on it can be made available to the scientist. However, many problems must be solved before protein databases become of general use to the scientific community. A most urgent one is to promote standardization of the gel running conditions so that data produced in a given laboratory may be used worldwide. Surprisingly, the gel running technology as it stands today is still a craftsmanship art.

Finally, comprehensive, computerized databases of proteins, together with recently developed techniques to microsequence proteins, offer a new dimension to the study of genome organization and function (Fig. 1). In particular, human protein databases may become increasingly important in view of the concerted effort to map and sequence the entire human genome. This formidable task is expected to dominate biological research in the next decades. [F]

We would like to thank S. Himmelstrup Jørgensen for typing the manuscript and O. Sønderkov for photography. Work in the authors' laboratories was supported by grants from the Danish Biotechnology Programme, the Danish Cancer Foundation, and the Commission of the European Communities.

## REFERENCES

- O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007-4021
- Special Issue: Two-dimensional gel electrophoresis. *Clin. Chem.* **28**, 1982
- Celis, J. E., and Bravo, R., eds. (1984) *Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications*. Academic, New York
- Celis, J. E., Madsen, P., Gesser, B., Kwee, S., Nielsen, H. V., Rasmussen, H. H., Honoré, B., Leffers, H., Ratz, G. P., Basse, B., Lauridsen, J. B., and Celis, A. (1989) Protein databases derived from the analysis of two-dimensional gels. In *Advances in Electrophoresis* (Chrambach, C., ed) VCH, Weinheim, Germany
- Special Issue: Two-dimensional gel electrophoresis in cell biology. (Celis, J. E., ed) *Electrophoresis* **11**, 1990
- Celis, J. E., Honoré, B., Bauw, G., and Vandekerckhove, J. (1990) Comprehensive computerized 2D gel protein databases offer a global approach to the study of the mammalian cell. *BioEssays* **12**, 93-98
- Garrels, J. I. (1983) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by cloned cell lines. *Methods Enzymol.* **100**, 411-423
- Anderson, N. L., Hofmann, J. P., Gemmel, A., and Taylor, S. (1984) Global approaches to the quantitative analysis of gene-expression patterns observed by two-dimensional gel electrophoresis. *Clin. Chem.* **30**, 2031-2036
- Garrels, J. I., Farrar, J. T., and Burwell, C. B. (1984) The Quest system for computer-analyzed two-dimensional electrophoresis of proteins in *Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications* (Celis, J. E., and Bravo, R., eds) pp. 37-91. Academic, New York
- Vincens, P., and Tarroux, P. (1988) Two-dimensional electrophoresis computerized processing. *Int. J. Biochem.* **20**, 499-509
- Appel, R., Hochstrasser, D., Roch, C., Funk, M., Müller, A. F., and Pellegrini, C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* **9**, 136-142
- Lemkin, P. F., and Lester, E. P. (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis* **10**, 122-140
- Miller, M. J. (1989) Computer-assisted analysis of two-dimensional gel electrophoretograms. *Adv. Electrophoresis* **3**, 182-217
- Phillips, T. D., Vaughn, V., Bloch, P. L., and Neidhardt, F. C. (1987) In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology, Gene-Protein Index of Escherichia coli K-12*, 2 ed. (Neidhardt, F. C., Ingraham, J. I., Low, K. B., Magasanik, B., Schaechter, M., and Umberger, H. E. ed) pp. 919-966. American Society for Microbiology, Washington, D.C.
- Celis, J. E., Ratz, G. P., Celis, A., Madsen, P., Gesser, B., Kwee, S., Madsen, P. S., Nielsen, H. V., Yde, H., Lauridsen, J. B., and Basse, B. (1988) Towards establishing comprehensive databases of cellular proteins from transformed human epithelial amnion cells (AMA) and normal peripheral blood mononuclear cells. *Leukemia* **9**, 561-601
- Special Issue: Protein databases in two-dimensional electrophoresis. (Celis, J. E., ed) *Electrophoresis* **2**, 1989
- Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Brogaard-Hansen, K. P., Kwee, S., Rasmussen, H. H., Nielsen, H. V., Crüger, D., Basse, B., Leffers, H., Honoré, B., Möller, O., and Celis, A. (1989) Computerized, comprehensive databases of cellular and secreted proteins from normal human embryonic lung MRC-5 fibroblasts: identification of transformation and/or proliferation sensitive proteins. *Electrophoresis* **10**, 76-115
- Garrels, J. I., and Franza, B. R. (1989) The REF52 protein database. Methods of database construction and analysis using the Quest system and characterizations of protein patterns from proliferating and quiescent REF52 cells. *J. Biol. Chem.* **264**, 5283-5298
- Celis, J. E., Crüger, D., Kiil, J., Dejgaard, K., Lauridsen, J. B., Ratz, G. P., Basse, B., Celis, A., Rasmussen, H. H., Bauw, G., and Vandekerckhove, J. (1990) A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis* **11**, 242-254
- Celis, J. E., Gesser, B., Rasmussen, H. H., Madsen, P., Leffers, H., Dejgaard, K., Honoré, B., Olsen, E., Ratz, G., Lauridsen, J. B., Basse, B., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) Comprehensive two-dimensional gel protein databases offer a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis* **12**, 989-1071



21. Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honoré, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B., Ratz, G., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) The MRC-5 human embryonal lung fibroblast two-dimensional gel cellular protein database: quantitative identification of polypeptides whose relative abundance differs between quiescent, proliferating and SV40 transformed cells. *Electrophoresis* 12, 1072-1113
22. Garrels, J. I., Franza, B. R., Chang, C., and Latter, G. (1990) Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis* 12, 1114-1130
23. Van Bogelen, R. A., Hutton, M. E., and Neidhardt, F. C. (1990) Gene-protein database of *Escherichia coli* K-12, 3rd ed. *Electrophoresis* 12, 1131-1166
24. Hewick, R. M., Hunkapiller, M. W., Hood, L. E., and Dreyer, W. J. (1981) A gas-liquid solid phase peptide and protein sequencer. *J. Biol. Chem.* 256, 7990-7997
25. Weber, K., and Osborn, M. (1985) In *The Proteins and Sodium Dodecyl Sulfate: Molecular Weight Determination on Polyacrylamide Gels and Related Procedures* (Neurath, H. et al., eds) Vol. 1, pp. 179-223, Academic, New York
26. Hunkapiller, M. W., Lujan, E., Ostrander, F., and Hood, L. E. (1983) Isolation of microgram quantities of proteins from polyacrylamide gels for amino acid sequence analysis. *Methods Enzymol.* 91, 227-236
27. Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., and Van Montagu, M. (1985) Protein-blotting on polybrene-coated glass-fiber sheets. *Eur. J. Biochem.* 152, 9-19
28. Aebersold, R. H., Teplow, D. B., Hood, L. E., and Kent, S. B. H. (1986) Electroblooming onto activated glass. *J. Biol. Chem.* 261, 4229-4238
29. Bauw, G., De Loose, M., Inzé, D., Van Montagu, M., and Vandekerckhove, J. (1987) Alterations in the phenotype of plant cells studied by NH<sub>2</sub>-terminal amino acid-sequence analysis of proteins electroblotted from two-dimensional gel-separated total extracts. *Proc. Natl. Acad. Sci. USA* 84, 4806-4810
30. Matsudaira, P. (1987) Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *J. Biol. Chem.* 262, 10035-10038
31. Eckerskorn, C., Mewes, W., Goretzki, H., and Lottspeich, F. (1985) A new siliconized-glass fiber as support for protein-chemical analysis of electroblotted proteins. *Eur. J. Biochem.* 176, 509-519
32. Moos, M., Jr., Nguyen, N. Y., and Liu, T.-Y. (1988) Reproducible high yield sequencing of proteins electrophoretically separated and transferred to an inert support. *J. Biol. Chem.* 263, 6003-6008
33. Kennedy, T. E., Gawinowicz, M. A., Barzilai, A., Kandel, E. R., and Sweatt, J. D. (1988) Sequencing of proteins from two-dimensional gels by using in situ digestion and transfer of peptides to polyvinylidene difluoride membranes: application to protein associated with sensitization in *Aplysia*. *Proc. Natl. Acad. Sci. USA* 85, 7008-7012
34. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987) Internal amino acid sequence analysis of protein separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* 84, 6970-6974
35. Bauw, G., Van Den Bulcke, M., Van Damme, J., Puype, M., Van Montagu, M., and Vandekerckhove, J. (1988) Protein electroblotting on polybase-coated glassfiber and polyvinylidene difluoride membranes: an evaluation. *J. Prot. Chem.* 7, 194-196
36. Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Leffers, H., Rasmussen, H. H., Nielsen, H. V., Crüger, D., Basse, B., Honoré, B., Møller, O., Celis, A., Vandekerckhove, J., Bauw, G., Van Damme, J., Puype, M., and Van Den Bulcke, M. (1989) Comprehensive, human cellular protein databases and their implication for the study of genome organization and function. *FEBS Lett.* 244, 247-254
37. Bauw, G., Van Damme, J., Puype, M., Vandekerckhove, J., Gesser, B., Lauridsen, J. B., Ratz, G. P., and Celis, J. E. (1989) Protein-electroblotting and -microsequencing strategies in generating protein databases from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* 86, 7701-7705
38. Aebersold, R., and Leavitt, J. (1990) Sequence analysis of proteins separated by polyacrylamide gel electrophoresis. Towards an integrated protein database. *Electrophoresis* 11, 517-527
39. Bauw, G., Rasmussen, H. H., Van Den Bulcke, M., Van Damme, J., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1990) Two-dimensional gel electrophoresis, protein electroblotting and microsequencing: a direct link between proteins and genes. *Electrophoresis* 11, 528-536
40. Tempst, P., Link, A. J., Riviere, L. R., Fleming, M., and Ellicone, C. (1990) Internal sequence analysis of protein separated on polyacrylamide gels at the submicrogram level: improved methods, applications and gene cloning strategies. *Electrophoresis* 11, 537-553
41. Eckerskorn, C., and Lottspeich, F. (1990) Combination of two-dimensional gel electrophoresis with microsequence and amino acid composition analysis: improvement of speed and sensitivity in protein characterization. *Electrophoresis* 11, 554-561
42. Rasmussen, H. H., Van Damme, J., Bauw, G., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1991) In *Methods in Protein Sequence Analysis* (Jörnvall, H., and Höög, J. O., eds) pp. 103-114. Eighth International Conference on Methods in Protein Sequence Analysis. Birkhäuser Verlag, Boston
43. Olson, A. D., and Miller, M. J. (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem.* 169, 49-70
44. Vincens, P., Paris, N., Pujol, J. L., Gaboriaud, C., Rabilloud, T., Pennetier, J., Matherat, P., and Tarroux, P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* 7, 347-356
45. Celis, J. E., Madsen, P., Celis, A., Nielsen, H. V., and Gesser, B. (1987) Cyclin (PCNA, auxiliary protein of DNA polymerase- $\delta$ ) is a central component of the pathway(s) leading to DNA replication and cell division. *FEBS Lett.* 220, 1-7
46. Anderson, N. G., and Anderson, N. L. (1982) The human protein index. *Clin. Chem.* 28, 739-748



Bo Franzén<sup>1</sup>  
 Stig Linder<sup>2</sup>  
 Ken Okuzawa<sup>2</sup>  
 Harabumi Kato<sup>2</sup>  
 Gert Auer<sup>1</sup>

## Nonenzymatic extraction of cells from clinical tumor material for analysis of gene expression by two-dimensional polyacrylamide gel electrophoresis

<sup>1</sup>Division of Tumor Pathology,  
 Department of Pathology, Division  
 of Experimental Oncology,  
 Karolinska Hospital and Institute,  
 Stockholm Sweden

<sup>2</sup>Tokyo Medical College, Department  
 of Surgery, Tokyo

<sup>3</sup>Division of Experimental Oncology,  
 Karolinska Hospital and Institute,  
 Stockholm

We have compared different methods of preparation of malignant cells for two-dimensional electrophoresis (2-DE). We found all methods using fresh tissue to be superior compared to methods using frozen tissue. Our results indicate that nonenzymatic methods of preparation of tumor cells, including fine needle aspiration, scraping and squeezing, have advantages over methods using enzymatic extraction of cells. Nonenzymatic methods are rapid, appear to reduce loss of high molecular protein species, and alleviate the necessity of separating viable and nonviable cells by Percoll gradient centrifugation. Using these techniques, high-quality 2-DE maps were derived from tumors of the lung and breast. In the resulting polypeptide patterns, heat shock proteins, non-muscle tropomyosins and intermediate filament were identified. We conclude that nonenzymatic extraction of malignant cells from fresh tumor tissue improves the possibilities that these techniques may be useful in clinical diagnosis.

### 1 Introduction

Tumors may develop by a number of different mechanisms in any given cell type. At the time of diagnosis, tumors will have progressed along different pathways to various stages of malignancy. To provide a basis for individual therapy it is of importance to examine specific properties of the tumor cell population in each patient. A large number of different markers have been described in order to increase the diagnostic accuracy. It is likely that a combination of several markers is needed in the future in order to reflect different properties of the tumor. One important method for the resolution of a large number of potential markers is two-dimensional electrophoresis (2-DE). Extensive efforts are being made in identifying various polypeptides separated by 2-DE and to characterize how the expression of these polypeptides is affected by the response to cellular transformation and various culture conditions [1,2]. It would be of value to transfer this information to 2-DE separations of polypeptides from tumor tissue samples. However, one prerequisite is that the quality of the 2-DE gels from tumor samples is comparable in quality with 2-DE gels from samples of cultured cells.

Frozen tumor tissues are commonly used for various biochemical assessments. However, if such samples are analyzed by 2-D polyacrylamide gel electrophoresis (PAGE), the polypeptide patterns are obscured by contamination of serum- and connective tissue proteins. Such nontumor-cell-related variations represent serious problems in the interpretation and inter-patient comparison of 2-DE

patterns [3]. 2-DE patterns of cells prepared from fresh tumor material were analyzed after enzymatic extraction of tumor cells [4, 5] or after culturing tumor fragments in medium containing radioactive amino acids [6]. These procedures may, however, lead to alterations in the gene expression/polypeptide patterns. We are only aware of one study where nonenzymatic extraction of cells from fresh tumor tissue (prostate cancer) was used to prepare samples for 2-D PAGE [4]. We have examined enzymatic extraction and various nonenzymatic preparation techniques, including fine needle aspiration, for the preparation of cells from fresh tumor tissues. We describe nonenzymatic extraction procedures that are rapid, lead to high-quality 2-DE patterns, and that alleviate the necessity to purify tumor cell populations from dead cells.

### 2 Materials and methods

#### 2.1 Cell cultures and samples used for spot identification

A rat embryonal fibroblast cell line, WT2 (a kind gift from Dr. J. I. Garrels and Dr. S. Pattersson) was used for the identification of a number of heat shock and structural proteins. Human normal diploid lung fibroblasts, WI38, human epithelial breast carcinoma cells, MDA-231 and MCF-7 were purchased from ATCC and grown as recommended. Polypeptides prepared from a leukemia type pre-B-ALL were separated by 2-DE. The 2-DE map was then analyzed by Dr. S. M. Hanash (University of Michigan, Ann Arbor, USA).

#### 2.2 Tumor tissues samples

In this study, 2-DE maps from seven tumors were used as representative illustrations: two adenocarcinoma of the lung (LA, and LB, mucinous, both cases intermediate grade of differentiation), one squamous carcinoma of the lung (LS), one carcinoid-like breast cancer (BC), one microfollicular adenoma (highly differentiated) of the thyroid (TA), one highly differentiated hyperneph-

Correspondence: Dr. Bo Franzén, Division of Tumor Pathology, Department of Pathology, L1:01, Karolinska Hospital and Institute, 10401 Stockholm 60, Sweden

Abbreviations: 2-DE, Two-dimensional polyacrylamide gel electrophoresis; IEF, isoelectric focusing; LDH, lactate dehydrogenase; NP-40, Nonidet P-40; PBS, phosphate buffered saline; PCNA, proliferating cell nuclear antigen; PIH, protease inhibitors; PMSF, phenylmethyl sulfonyl fluoride; SDS, sodium dodecyl sulfate; WW, wet weight

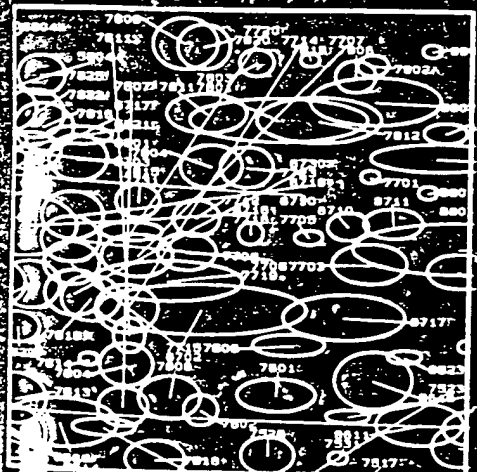
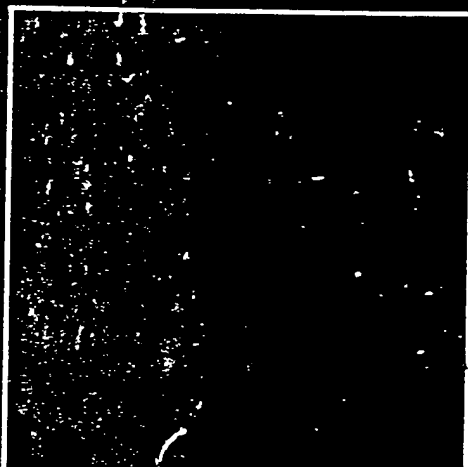
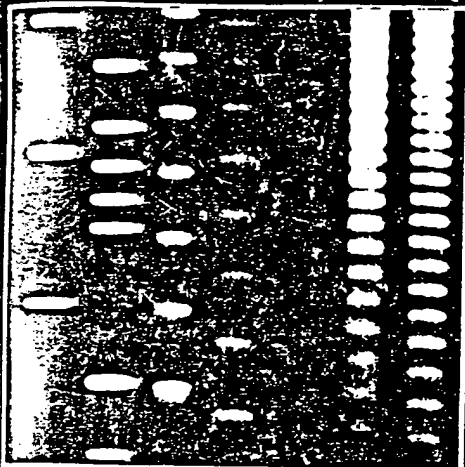
IS  
RS  
35  
(3)

SWAN

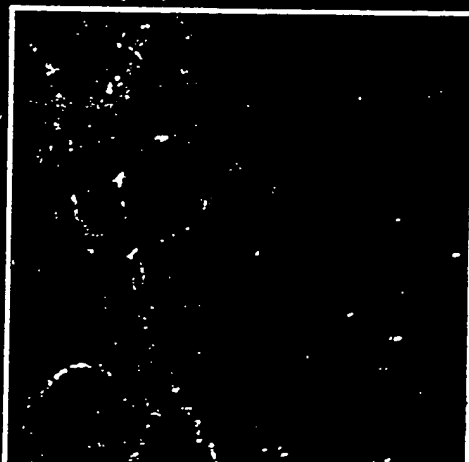
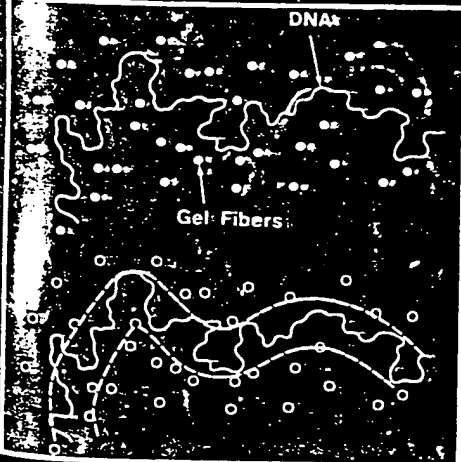
# ELECTROPHORESIS

An International Journal

10'93



**Applications  
Methods  
Theory**



roma, a tumor of the kidney (KH), and finally one case of poorly differentiated corpus carcinoma (CP).

### 2.3 Preparation of cultured cells

The cell monolayers were washed twice in phosphate buffered saline (PBS) and then scraped off in ice-cold PBS including protease inhibitors (PIH), phenylmethylsulfonyl fluoride (PMSF) 0.2 mM and 0.83 mM benzamide pelleted at  $660 \times g$ , 3 min ( $+4^{\circ}\text{C}$ ) and washed one time before final centrifugation at  $2700 \times g$ , 5 min. The wet weight of the cell pellet was recorded and the cells were stored at  $-80^{\circ}\text{C}$  until further processing.

### 2.4 Preparation of tumor tissue samples

#### 2.4.1 General remarks

Macroscopically representative and non-necrotic tumor tissues were selected within 20 min after resection. Parallel samples were routinely prepared for cytology. The samples were processed as rapidly as possible on ice or at  $+4^{\circ}\text{C}$  and in the presence of PIH. Cells were stained with DiffQuick (Baxter) and usually examined at three different occasions during the preparation procedure: (i) cytology sample, (ii) extracted cells and (iii) cells after percoll gradient centrifugation.

#### 2.4.2 Specimen acquisition

The strategy of sample preparation is shown in Fig. 1. Tumor tissue cell samples were usually obtained by fine needle aspiration (NA) using a 0.7 mm needle. The syringe was filled with 1–2 mL of ice-cold culture medium/PIH. We found that if a tumor appeared to be very fibrous it is difficult to extract enough cells for 2-DE analysis. In these cases, two alternative techniques were examined. (i) The tumor was cut in the middle and the fresh surface scraped (SC) by a scalpel. The cell-rich material was then transferred to ice-cold culture medium (L15 with 5% fetal calf serum)/PIH. (ii) A part of the tumor sample was placed in culture medium on ice for further processing at the laboratory in the following way: the material was cut into very small fragments on a pre-cooled dissection plate and transferred to a small glass chamber with a 0.7 mm metal net 5 mm above the bottom of the chamber. Medium /PIH was added to cover the sample (8 mL) which was gently squeezed (SQ) towards the net in order to release and wash out cells. NA and SC were also compared with an enzymatic extraction (EE) procedure described previously [5]. Briefly, thin slices of tissue were incubated with collagenase (1 mg/mL) and elastase (2 mg/mL) in medium for 1 h at  $37^{\circ}\text{C}$ . Extracted cells from every sample were then subjected to percoll gradient centrifugation (Section 3.2.3).

#### 2.4.3 Separation of cells by Percoll gradient centrifugation

The cell suspension was filtered through two nylon mesh filters, (i) 250  $\mu\text{m}$  and (ii) 100  $\mu\text{m}$  and then centrifuged

at  $660 \times g$  for 3 min. The cell pellet was resuspended carefully in medium, using a syringe and loaded onto a two-step discontinuous Percoll/PBS gradient, 20.4 (density = 1.03 g/mL) and 54.7% (density = 1.07 g/mL), and centrifuged at  $1000 \times g$  for 15 min. In this system, dead cells stay on the top, viable cells sediment to the interphase and erythrocytes sediment to the bottom. The viability of cells in the top fraction and interphase was checked by the trypan blue exclusion test. The interphase cell layer ( $> 90\%$  viability) was collected and washed one time in a large volume PBS/PIH (centrifuged at  $800 \times g$  for 3 min). Finally, the cells were resuspended in 1.4 mL PBS and pelleted at  $2700 \times g$  for 5 min. The wet weight (WW) was recorded and the pellet was then stored at  $-80^{\circ}\text{C}$ .

#### 2.4.4 Final preparation of cells for 2-D PAGE analysis

From this point, cultured cell samples were treated in the same way as tumor cell samples: Each cell pellet was thawed on ice and resuspended in 1.89  $\mu\text{L}$  MQ water per mg WW ( $= 1.89 \times \text{WW}$ )  $\mu\text{L}$ . The suspension was frozen and thawed 4–5  $\times$  to break the cells [7]. A volume of  $(0.089 \times \text{WW})$   $\mu\text{L}$  10% sodium dodecyl sulfate (SDS), including 33.3% mercaptoethanol, was mixed with the sample and incubated 5 min on ice with  $(0.329 \times \text{WW})$   $\mu\text{L}$  of a solution of DNase I (0.144 mg/mL 20 mM Tris-HCl with 2 mM  $\text{CaCl}_2 \times 2\text{H}_2\text{O}$ , pH 8.8) and RNase A (0.0718 mg/mL Tris) [8,9]. The sample was frozen and lyophilized. Sample buffer [10] including

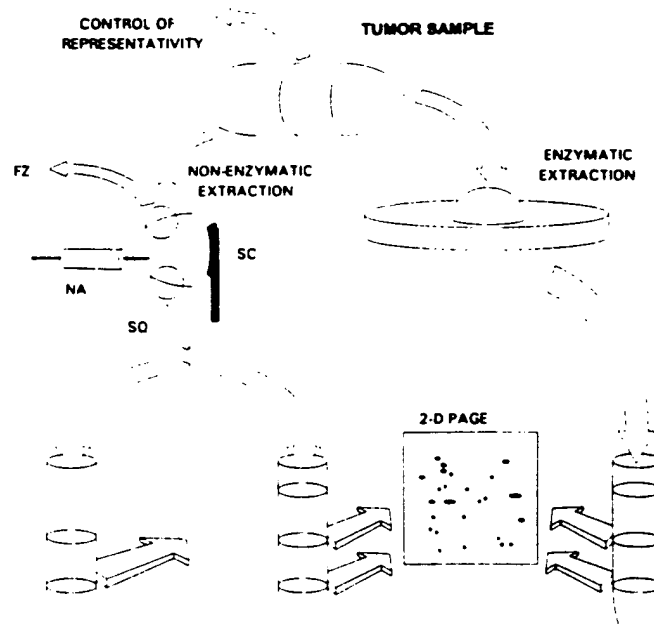


Figure 1. Experimental flow chart showing main steps of the preparation procedures. The abbreviations used for nonenzymatic extraction procedures are: FZ: frozen sample preparation; NA: needle aspiration; SC: scraped; and SQ: squeezed sample. Extracted cells are then loaded as a suspension (top volume of each tube) onto either 1.07 g/mL Percoll (left), or a discontinuous Percoll gradient from the nonenzymatic extraction (middle), or from enzymatic extraction (right). Cellular top- and interphase fractions are then used for 2-DE. For details see Section 2.

PMSF (0.2 mM), EDTA (1.0 mM), 0.5% Nonidet P-40 (NP-40), and 3-[3-cholamido propyl)-dimethylammonio]-1-propane sulfonate (CHAPS; 25 mM) was added carefully, mixed for 2.5 h and centrifuged for 15 min at

10000 rpm to remove any insoluble material. Duplicate or triplicate samples were taken for protein determination [11]. Samples were stored at  $-80^{\circ}\text{C}$  prior to isoelectric focusing (IEF).

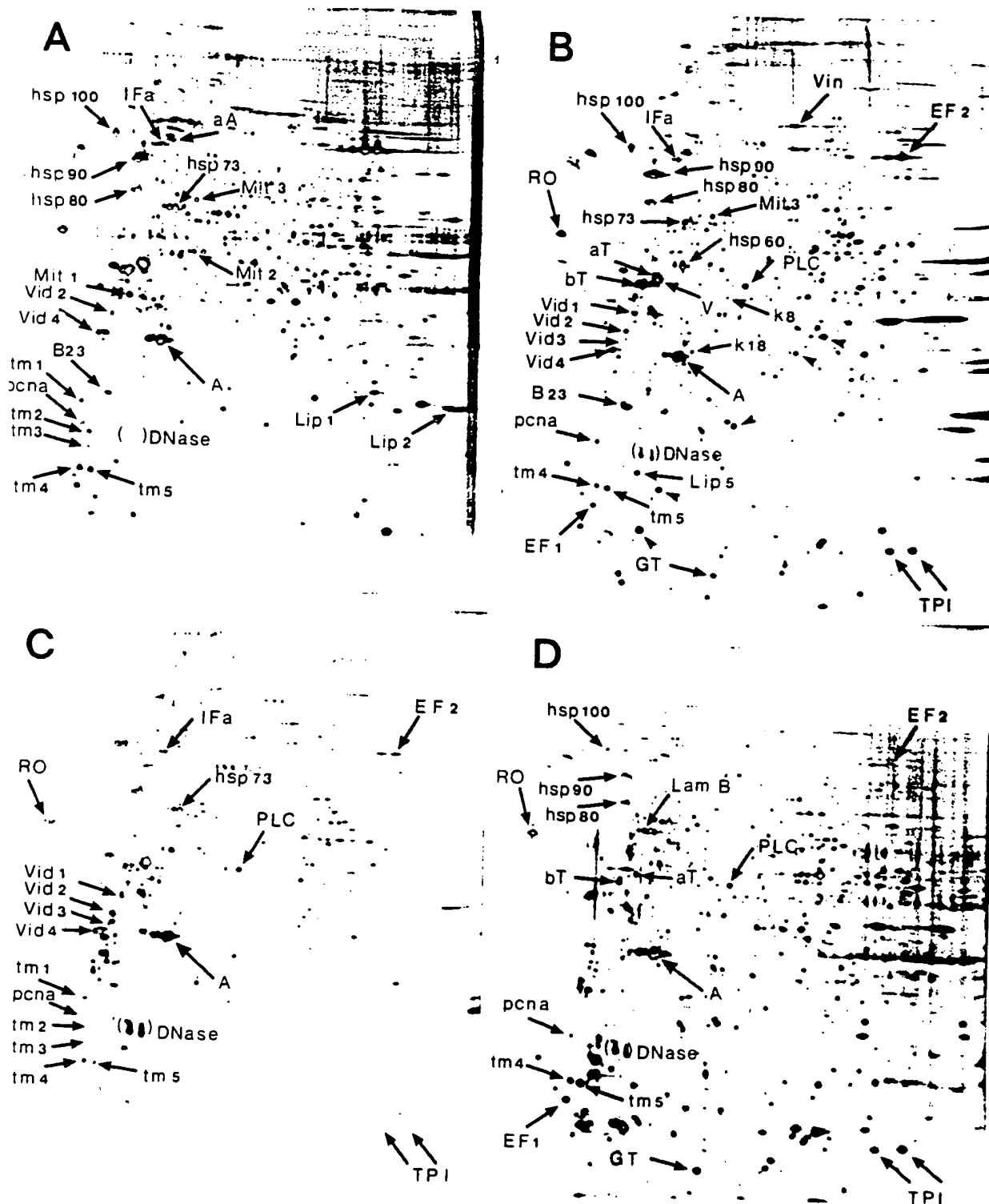


Figure 2. 2-DE analysis of samples from three cell lines and one leukemia used for the identification of polypeptides: (A) WT2; (B) MDA-231, arrowheads mark some low molecular weight cytosolic polypeptides; (C) W138 and (D) pre B-ALL. The abbreviations for identified spots are explained in Table 1.

### 2.4.5 Preparation of frozen tumor tissue

The technique has been described previously [3,12]. Briefly, the sample is moarted frozen to a fine powder, homogenized, lyophilized and solubilized in sample buffer.

### 2.4.6 Control of representativity

The tumors were examined routinely by experienced pathologists and smears or imprints from the samples were also assessed for cytometric DNA content by microspectrophotometry.

## 2.5 2-D PAGE

2-D PAGE was performed as described [8,10] except for the following details. The glass tubes for IEF,  $1.2 \times 200$  mm, contained 2.0% Resolyte, pH 4–8 (BDH) and were cast to a height of 180 mm. A stock solution of acrylamide (Serva) and *N,N'*-methylenebisacrylamide (16.7:1 for IEF and 37.5:1 for the second dimension) was deionized by mixing with 5% w/v Duolite MB 5313 mixed-resin ion exchanger (BDH) for 30 min, filtered (with a  $0.22 \mu\text{m}$  nitrocellulose filter) and stored at  $-70^\circ\text{C}$ . *N,N'*-Methylenebisacrylamide, *N,N,N',N'*-tetramethylethylenediamine (TEMED) and ammonium persulfate were purchased from Bio-Rad. IEF tubes were prefocused at 200 V in 60 min. To each tube a sample corresponding to 20–40  $\mu\text{g}$  protein was applied and focused for 14.5 h at 800 V and finally 1.0 h at 1000 V using a Protean II cell (Bio-Rad) and Model 1000/500 Power Supply (Bio-Rad). The tube gels were finally extruded into 1.25 mL equilibration buffer, containing 60 mM Tris, pH 6.8 (2% SDS, 100 mM dithiothreitol and 10% glycerol), frozen on dry ice and stored at  $-70^\circ\text{C}$ . The second dimension ( $1.0 \times 180 \times 90$  mm) of the acrylamide concentration was 10%

T, and the gel contained 376 mM Tris, pH 8.8, and 0.1% SDS. IEF gels were applied on top of the slab gel, sealed with 0.5% agarose containing electrophoresis running buffer (60 mM Tris-base, 0.2 M glycine and 0.1% SDS) and electrophoresed with 10–11 mA per gel (constant current) at  $+10^\circ\text{C}$ . Six gels were run together in a Protean II xi 2-D Multi-Cell (Bio-Rad). Proteins were visualized by silver staining and photographed with the acidic side to the left [13,14].

## 2.6 Identification of polypeptides

Vimentin and vimentin-derived polypeptides were identified by extraction of an MDA-231 cell lysate with 0.6 M KCl/0.5% NP-40 [15]. Tropomyosins were extracted from MDA-231 and WI38 cell lysates [16], and cytokeratins were extracted from MDA-231 and MCF-7 cell lysates [17]. The patterns were compared with published maps [19–21]. Proliferating cell nuclear antigen (PCNA) was identified by immunoblotting (PC10 mAb, Dako-patt) using a semidry system (Multiphor II Nova Blot, Pharmacia-LKB Biotechnology AB) and enhanced chemoluminescence (ECL) detection (Amersham).

## 3 Results

### 3.1 2-DE of samples prepared from normal and tumorigenic cultured cells

The object of this study was to develop methods for preparation of 2-DE maps from human tumor tissue which have the same high resolution as those obtained from cultured cells. Shown in Fig. 2 are high resolution 2-DE gels prepared from cultured cells and one leukemia: SV40 transformed embryonal rat fibroblasts WT2 (Fig. 2a); human MDA-231 breast carcinoma cells (Fig. 2b); human WI38 fibroblasts (Fig. 2c) and human pre B-ALL

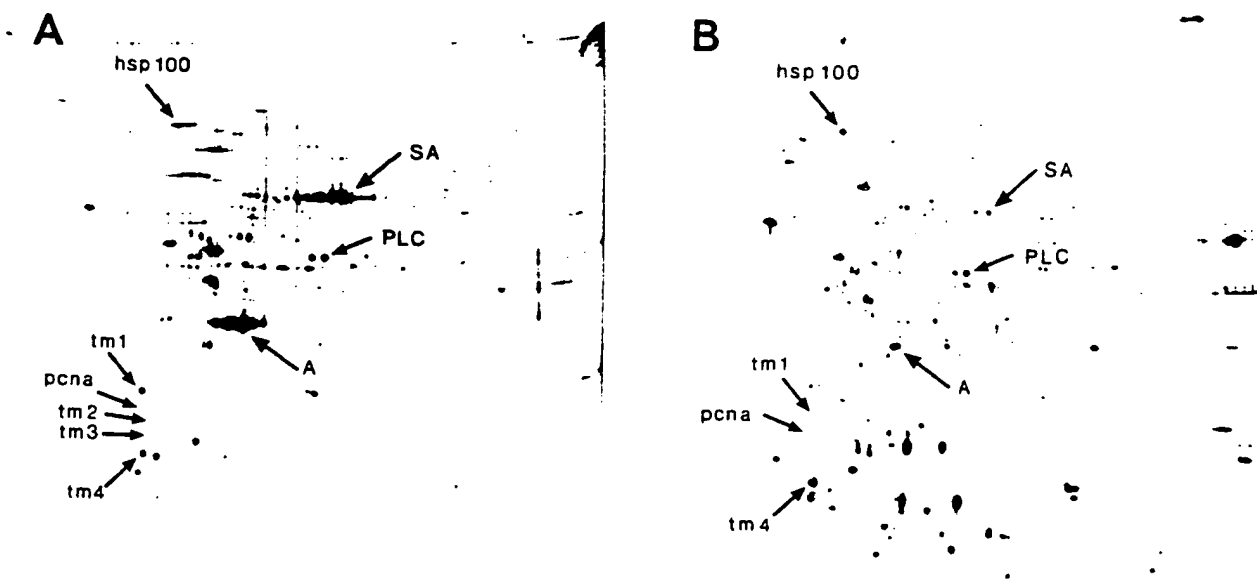


Figure 3. 2-DE analysis of a case of lung adenocarcinoma (LA). Comparison of 2-DE gel quality between (A) frozen and (B) fresh (needle aspiration) tissue preparation.

cells (Fig. 2d). Polypeptides were identified through a laboratory exchange of cell samples/2-DE maps and through 2-DE analysis of purified proteins (Table 1).

### 3.2 Preparation of samples from solid tumors

#### 3.2.1 Fresh versus frozen tissue

An adenocarcinoma of the lung (LA) was prepared for 2-DE by conventional methods using frozen material (Fig. 3a). There are several possibilities for the poor resolution using frozen tissue, including the presence of high molecular weight protein aggregates. Filtering extracts through 0.1  $\mu$ m filters (Durapore, Millipore) resulted in a slightly improved resolution (not shown). When fresh tumor tissue from tumor LA was used for sample preparation, using fine needle aspiration to collect the cells, the resolution was considerably improved (Fig. 3b). The use of fresh tissue resulted in a general increase in resolution, which was most pronounced in the 50–100 kDa molecular mass range. A number of differences in the protein profiles of the gels in Figs. 3a and 3b can be observed, some of which are indicated in the figures. The decrease in serum albumin in Fig. 3b is likely to result from loss of serum proteins occurring when cells were pelleted after aspiration. Other differences, such as the decreased level of transformation-sensitive tropomyosins (TM1-TM3), may result from enrichment of tumor cells in the sample of Fig. 3b. Fine needle aspiration, a well-established technique in cytology, extracts mainly tumor cells because of decreased intercellular adhesiveness of neoplastic cells as compared to normal tissue. Microscopic examination of Diff-Quick-stained extracted cells from case LA revealed almost 100% tumor cells, whereas the whole tissue extract contained approximately 60% tumor cells.

Table 1. Names and abbreviations for identified spots

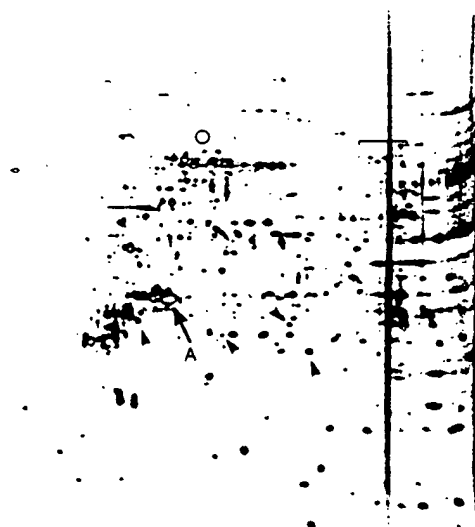
Spot	Name	Basis for identification
A	Actins	a
aA	$\alpha$ -Actinin	a
B23	Protein B23 /Numatrin	a
EF2	Elongation factor 2	a
EF1	Elongation factor 1 $\beta$	a
GT	Glutathione-S-transferase ( <i>pi</i> )	a
hsp60	Heat shock protein 60	a
hsp73	Heat shock protein 73	a
hsp80	Heat shock protein 80, GRP78, BIP	a
hsp90	Heat shock protein 90	a
hsp100	Heat shock protein 100, Endoplasmic	a
IFa	Intermediary filament associated	a
k8	Cytokeratin 8	b and a
Lamb	Lamin B	a
Lip1	Lipocortin I	a
Lip2	Lipocortin II	a
Lip5	Lipocortin V	a
Mit1	Mitcon 1/ $\beta$ - F1 ATPase	a
Mit2	Mitcon 2	a
Mit3	Mitcon 3	a
MRP	Mucine Related Polypeptides	—
pca	Proliferating cell nuclear antigen	c and a
PLC	Phospholipase C (1)	a
RO	RO/SS-A antigen	a
SA	Serum Albumin	b and a
aT	$\alpha$ -Tubulin	a
bT	$\beta$ -Tubulin	a
tm1	Non-muscle tropomyosin isoform 1	b and a
tm2	Non-muscle tropomyosin isoform 2	b and a
tm3	Non-muscle tropomyosin isoform 3	b and a
tm4	Non-muscle tropomyosin isoform 4	b and a
tm5	Non-muscle tropomyosin isoform 5	b and a
TPI	Triose phosphate isomerase	a
V	Vimentin	b and a
Vid1	Vimentin derived protein	b and a
Vid2	Vimentin derived protein	b and a
Vid3	Vimentin derived protein	b and a
Vid4	Vimentin derived protein	b and a
Vin	Vinculin	a

a. homologous position with respect to other mammalian systems

b. purified protein(s)

c. immunoblotting

A



B

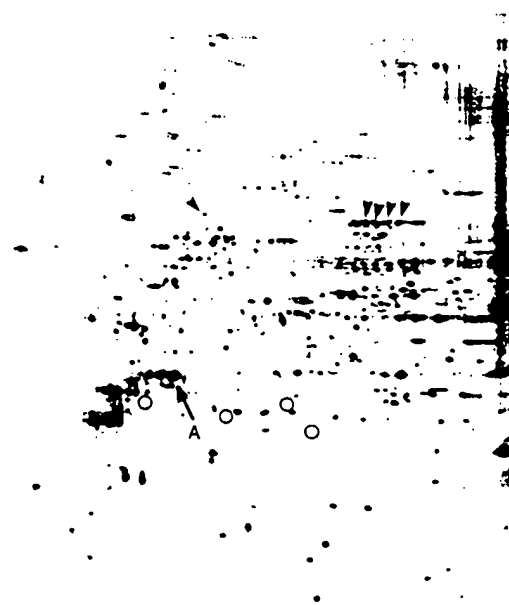


Figure 4. 2-DE analysis of a case of breast carcinoma (BC). Comparison of 2-DE quality and some differences in detected spots (arrow heads indicate increased intensity and circles or bracket indicate decreased intensity of the same spots) between (A) enzymatically and (B) nonenzymatically (scraped) tissue preparation.

### 3.2.2 Comparison of different methods for preparing cells from fresh tumor tissue

Samples were prepared from breast and lung carcinomas using either an enzymatic treatment with collagenase/elastase or using nonenzymatic preparations (Fig. 4). A number of differences in the protein profiles were observed in the resulting 2-DE gels, some of which are indicated in Figs. 4a and b. These differences include both increases and decreases in spot intensity. These differences may result from degradation of high molecular weight polypeptides during enzymatic treatment, increased solubilization of polypeptides, or may have other causes. For many tumors, it was only possible to obtain

small amounts of material since they were reserved for other examinations. In these cases, samples could be prepared for 2-DE using either needle aspiration or scraping. Figure 5a shows a 2-DE gel prepared from squamous lung carcinoma (LS) cells collected by needle aspiration and Fig. 5b shows a gel prepared from the same tumor by scraping. In this case, a number of differences were recorded between the two procedures, some of which are arrowed in Fig. 5. Samples obtained from other tumors (breast and lung) generally showed fewer differences between these two methods of cell sampling (not shown). These data show that different nonenzymatic extraction procedures may yield different polypeptide patterns. However, the number of spots with a large

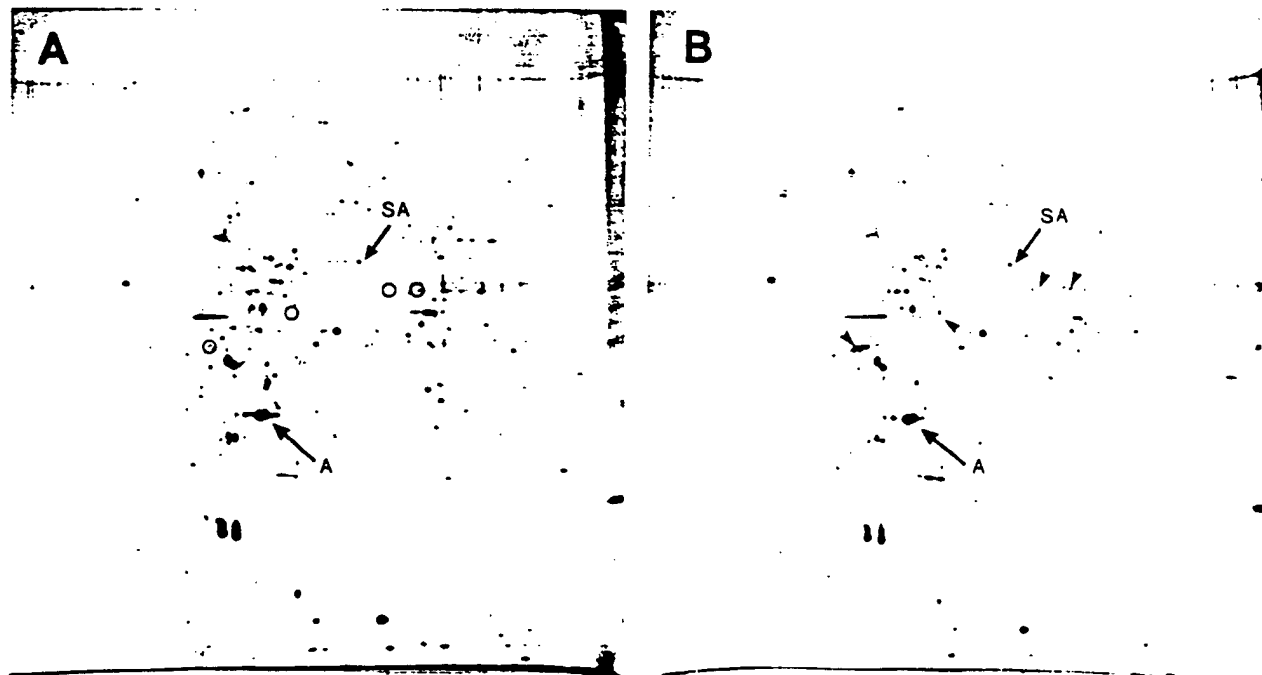


Figure 5. 2-DE analysis of a case of lung cancer (LS). Comparison of 2-DE gel quality and detected spots (arrow heads and circles) between (A) aspirated (needle aspiration) and (B) scraped preparations from fresh tissue.

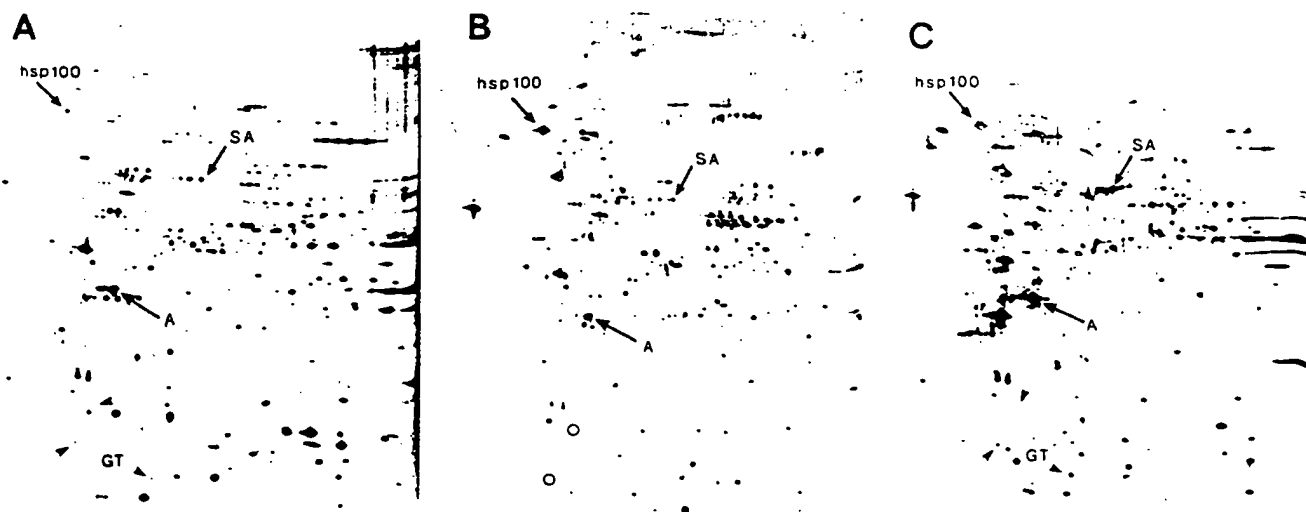


Figure 6. 2-DE analysis of three other types of tumors, (A) hypernephroma, (B) an adenoma of the thyroid and (C) corpus cancer, using the nonenzymatic preparation technique. Arrowheads and circles indicate some cytosolic polypeptides.

difference in intensity were lower than when a nonenzymatic preparation was compared with an enzymatic preparation.

2-DE maps of satisfactory quality were prepared by a third procedure. Cells were released from small pieces of tumor by squeezing (see Section 2). Some examples of this are shown in Fig. 6 where 2-DE maps derived from a case of hypernephroma, KH (Fig. 6a), a case of thyroid tumor, TA (Fig. 6b) and a case of corpus cancer, CP (Fig. 6c) can be seen. We conclude that nonenzymatic techniques are useful for 2-DE analysis of a number of different tumors. The quality of the resulting gels is com-

parable to that obtained using cultured cells (compare the gels in Fig. 2 with those in Fig. 4, 6 and 7). Which of these methods will be optimal will, in our experience, depend on the tumor material. For example, very small tumors are preferably extracted by squeezing; on the other hand, breast cancers (which are often fibrous) yield satisfactory samples using scraping.

### 3.2.3 Purification of cells on percoll gradients

We considered the possible advantage of separating viable cells from dead cells, erythrocytes, and debris using discontinuous Percoll gradients. Cells collected

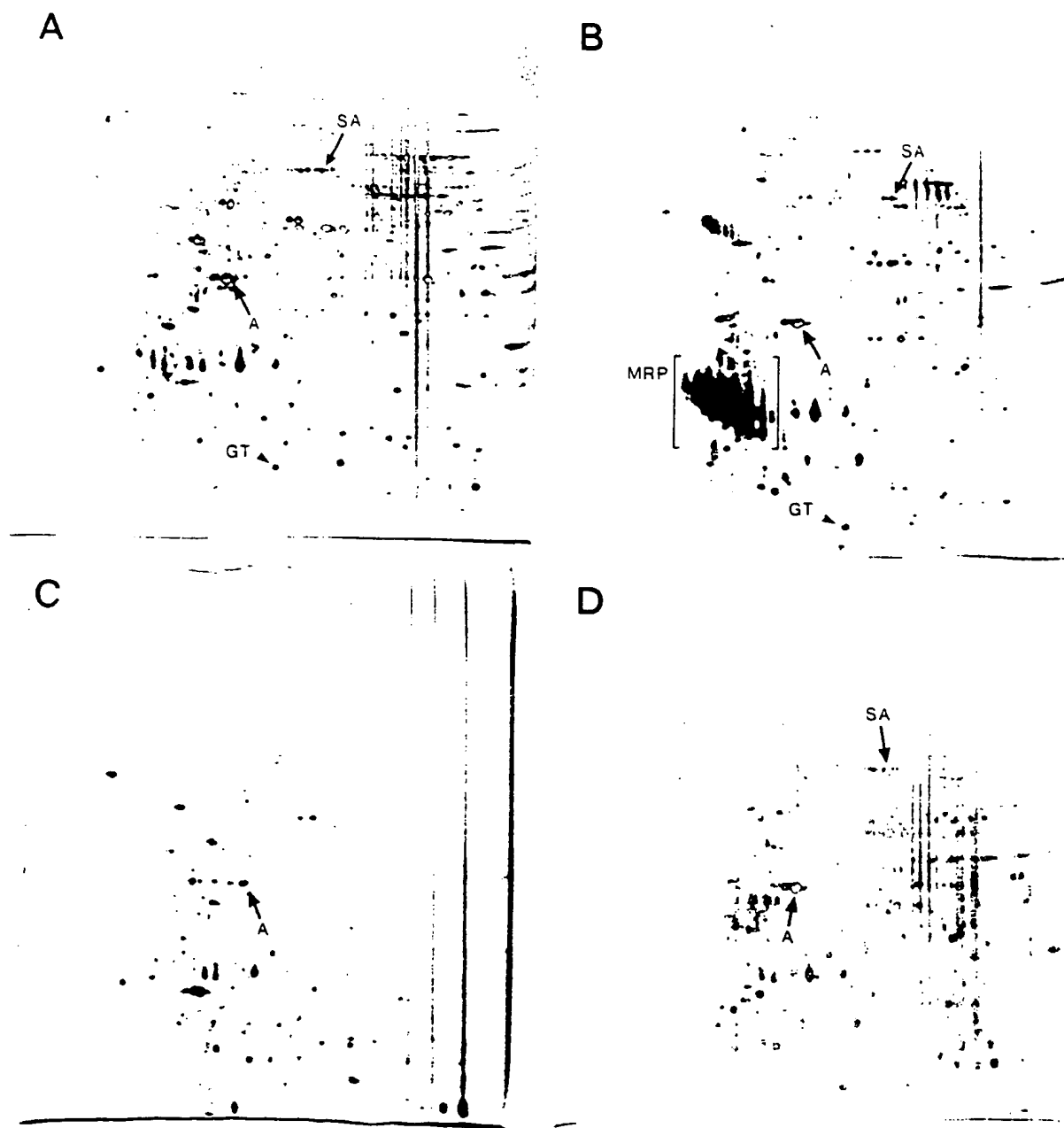


Figure 7. 2-DE analysis of polypeptides from viable (b and d) and nonviable (a and c) cells of an adenocarcinoma of the lung (LB), separated using discontinuous Percoll density gradient. Nonenzymatic preparation technique (a and b) and enzymatic preparation technique (c and d) are compared.



from the interphase showed a viability of more than 90% as judged by trypan blue exclusion test. However, it was found that the yield of viable cells decreased dramatically if the tissue resection was not immediately processed. To study the effect of lysis of cells during the preparation procedure, 2-DE maps were prepared from nonenzymatically extracted cells of case LB collected from the top fraction (nonviable, Fig. 7a) and interphase fraction (viable, Fig. 7b). These 2-DE maps were compared with corresponding fractions (nonviable, Fig. 7c, and viable, Fig. 7d) of enzymatically extracted cells. One clear disadvantage of the enzymatic technique was that when loss of cell viability occurred during preparation, a dramatic loss of high molecular weight polypeptides was observed (Fig. 7c). This was probably due to degradation of intracellular proteins. However, nonenzymatic preparations showed fewer differences between viable and nonviable cells: The most pronounced alteration was a decrease of a group of mucine related proteins (Fig. 7b). We conclude, therefore, that discontinuous Percoll gradient is necessary after enzymatic extraction of cells, but can be omitted from the nonenzymatic tumor sample preparation procedure.

We used the MDA-231 cell line to study the effects of cell lysis and leakage of cytosolic polypeptides during sample preparation. Remarkably, after 30, 50, 80 and 140 min of incubation in PBS/PIH at 0°C, no significant changes were observed in the 2-DE pattern (not shown). Although loss of cell viability may not result in protein degradation when cells are incubated in the presence of protease inhibitors, loss of cytosolic proteins would be expected during pelleting of cells. We monitored the loss of lactate dehydrogenase (LDH) activity into the supernatant during incubation in PBS of MDA-231 and MCF-7 breast cancer cells at 20°C. In both cases, loss of viability was paralleled by release of LDH from the cells (Fig. 8). After 5 h, 70% of the MCF-7 cells, but only 30% of the MDA-231 cells were dead (not shown).

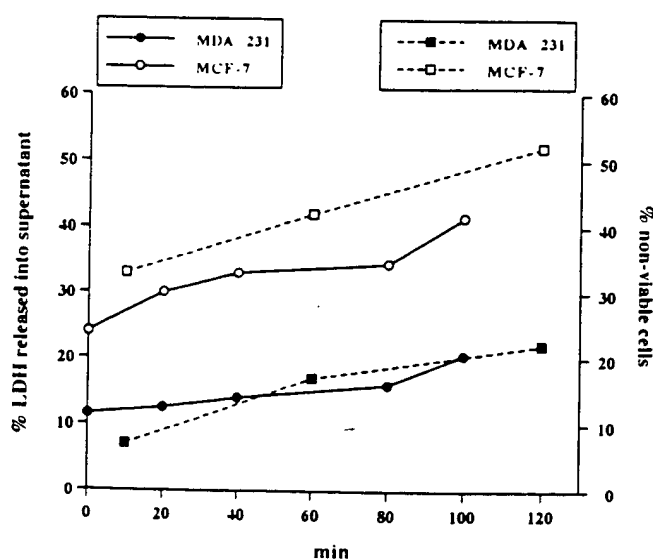


Figure 8. The relative release (fraction in supernatant of total) of lactate dehydrogenase activity (LDH) and cell viability versus incubation time of the mammary carcinoma cell lines MDA-231 and MCF-7 during incubation in PBS at 20°C.

These data indicate the impact of a rapid preparation procedure, at low temperature, of fresh tumor samples. Experiments have also been performed using only 1.07 g/mL Percoll (Fig. 6c and Fig. 1, left test tube) in order to remove erythrocytes. One clear advantage with this procedure, which today is routinely utilized, is a higher yield of viable cells, probably due to decreased sample preparation time.

#### 4 Discussion

We describe procedures for sample preparation from solid tumors for 2-DE. 2-DE maps could be derived from solid tumors which were similar in quality to those obtained from cultured cells. Compared to methods using frozen material, the resolving power of the 2-DE technique is increased, allowing examination of a large number of polypeptides from tumors of different malignancies. Other investigators [12,22] have used samples from frozen tumors to derive 2-DE maps. We have previously described disadvantages encountered using frozen tumor samples including variations in contaminating proteins between different samples [3]. The methods described here are based on the preparation of cells from tumors without enzymatic digestion. The enzymatic step could be avoided since malignant cells usually grow as solid masses which are not strongly attached to the matrix. Furthermore, we found that omitting the enzymatic digestion alleviated the necessity of purifying viable tumor cells on Percoll gradients. This was in sharp contrast to enzymatically treated samples, where loss of viability leads to loss of high molecular weight proteins (Fig. 7c).

At least in the case of lung cancer, viable and nonviable cells showed small differences in respect to 2-DE maps. Presumably, protease inhibitors penetrate cells and inhibit proteolysis. In model experiments, we observed leakage of cytosolic protein (LDH) from the cells in parallel to loss of viability. Apparently, however, only a limited decrease of the level of low molecular weight cytosolic polypeptides was detected using silver staining combined with visual inspection. We have found that although some tumors are well suited for the preparation procedure described, others are not. In general, good results were obtained using tumors of the lung, breast, corpus and lymphomas. In contrast, cells from thyroid adenomas and hypernephroma showed poor viability. We were in these cases unable to separate nonviable cells from viable cells, and we can therefore not evaluate the consequence of the loss of viability on 2-DE patterns, apart from a loss of some low molecular weight cytosolic polypeptides.

Highly differentiated tumors may show lower viability as compared with poorly differentiated tumors (Dr. Farkas Vanky, personal communication). A number of samples from thyroid tumors were prepared for 2-DE but most cases showed poor viability. We believe that special care is needed during preparation of generally highly differentiated tumor groups. The difference between loss of viability/leakage of LDH of the more differentiated MCF-7 cells and the less differentiated MDA-231 cells is in line

with these observations (Fig. 8). A number of potential and interesting markers, like tropomyosin isoforms, cytokeratins and heat shock proteins, appear to be insensitive to loss of viability during the preparation procedure. We have to date made numerous observations of alterations in the expression of these polypeptides in breast cancers and lung cancers.

Another problem that may occur, irrespective of sample preparation techniques used, is admixture of lymphocytes. These cases are easily detectable in smears and it may therefore be possible to select lymphocyte specific spots as "internal markers" for the 2-D PAGE analysis. Studies using this approach are in progress. Many of the polypeptides identified are structural (Table 1). Since the expression of many of these polypeptides are known to vary between normal and malignant cells, the possibility to determine their expression simultaneously is appealing. In the specific case of breast cancer, alterations in the expression of intermediate filament proteins (cytokeratins) are known to occur during tumor progression [23]. Other proteins known to be differentially expressed between normal cells and transformed cells are tropomyosins, numatrin/B23, heat shock proteins and PCNA. To this end, we have observed alterations in the expression of cytokeratin 8, hsp 90, and non-muscle tropomyosin isoform 2 during malignant progression. (Okuzawa *et al.*, in preparation and Franzén *et al.*, in preparation).

The method of choice for sample preparation from tumor tissues will depend on the properties of the tumor material studied. It may be important to use only one method when comparing cases within one group, as differences were observed between methods. The advantages of the nonenzymatic techniques are (i) that it minimizes contamination with connective tissue, (ii) that problems with contamination of serum proteins are avoided, and (iii) that separation of viable and dead cells is not necessary. Hereby the resolving power of 2-D PAGE is maximized for the analysis of human tumors and studies on inter-tumor variations in gene expression are facilitated. In addition, the polypeptide patterns obtained may be more representative for the *in vivo* tumor cell since the use of enzymes and incubations have been minimized.

We would like to thank Dr. J. I. Garrels, Dr. S. Patterson, Dr. S. M. Hanash and Dr. J. E. Celis for making sample and 2-DE map exchanges possible. This study was sup-

ported by grants from the Swedish Cancer Society and the Cancer Society in Stockholm.

Received March 5, 1993

## 5 References

- [1] Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honore, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B. and Ratz, G., *Electrophoresis* 1990, 11, 1072-1113.
- [2] Garrels, J. I., Franza, B. R., Chang, C., Latter, G., *Electrophoresis* 1990, 11, 1114-1130.
- [3] Franzén, B., Iwabuchi, H., Kato, H., Lindholm, J. and Auer, G., *Electrophoresis* 1991, 12, 509-515.
- [4] Sherwood, E. R., Berg, L. A., Mitchell, N. J., McNeal, J. E., Kozlowski, J. M. and Lee, C., *J. Urology* 1990, 143, 167-171.
- [5] Endler, A. T., Young, D. S., Wold, L. E., Lieber, M. M. and Curie, R. M., *J. Clin. Chem. Clin. Biochem.* 1986, 24, 981-992.
- [6] Forchhammer, J. and Macdonald-Bravo, H., in: Celis, J. E. and Bravo, R., (eds.), *Gene Expression in Normal and Transformed Cells*, Plenum, New York 1983, pp. 291-314.
- [7] Linder, S., Brzeski, H. and Ringertz, N. R., *Exp. Cell. Res.* 1979, 120, 1-14.
- [8] Celis, J. E. and Bravo, R. (Eds.), *Two-dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 3-36.
- [9] Garrels, J. I., *J. Biol. Chem.* 1979, 254, 7961-7977.
- [10] Anderson, N. L., *Two-Dimensional Electrophoresis, Operation of the ISO-DALT System*, Large Scale Biology Press, Washington, DC 1988, 162.
- [11] Bradford, M., *Anal. Biochem.* 1976, 72, 248.
- [12] Tracy, R. P., Wold, L. E., Currie, L. M. and Young, D. S., *Clin. Chem.* 1982, 28, 890-899.
- [13] Merrill, C. R., Goldman, D., Sedman, S. A. and Elbert, H. M., *Science* 1981, 211, 1437-1438.
- [14] Morrissey, J. H., *Anal. Biochem.* 1981, 117, 307-310.
- [15] Gard, D. L., Bell, P. B., Lazarides, E., *Proc. Natl. Acad. Sci. USA*, 1979, 76, 3894-3898.
- [16] Matsumura, F., Lin, J.-C., Yamashiko-Matsumura, S., Thomas, G. P. and Topp, W. C., *J. Biol. Chem.*, 1983, 258, 13954-13960.
- [17] Paulin, D., Forest, N. and Perreau, J., *J. Mol. Biol.* 1980, 144, 95-101.
- [18] Blobel, G. A., Moll, R., Franke, W. W., Kayser, K. W. and Gould, V. E., *Am. J. Pathol.* 1985, 121, 235-247.
- [19] Ochs, D. C., McConkey, H. E. and Guard, N. L., *Exp. Cell. Res.* 1981, 135, 355-362.
- [20] Bhattacharya, B., Gaddamanuga, L. P., Valverius, E. M., Salomon, D. S. and H. L. Cooper, *Cancer Res.* 1990, 50, 2105-2112.
- [21] Sommers, C. L., Walker-Jones, D., Heckford, S. E., Worland, P., Valverius, A., Clark, R., McCornick, F., Stampfer, M., Abularch, S. and Gelmann, E. P., *Cancer Res.* 1989, 49, 4258-4263.
- [22] Trask, D. K., Band, V., Zajchowski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, 87, 2319-2323.
- [23] Trask, D. K., Bond, V., Zajchowski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, 87, 2319-2323.

## Company Info

### LSB & LSP Information

Large Scale Biology Corporation

Large Scale Proteomics Corporation

---

#### Large Scale Biology Corporation

***Large Scale Biology Corporation is the leader in the integrated discovery, production and application of proteins - the functional units of all biological processes.***

Large Scale Biology Corporation (LSB, Vacaville, CA) and its subsidiary Large Scale Proteomics Corp. (LSP, Germantown, MD) are a biotechnology enterprise with the mission of accelerating the speed and productivity of the life sciences industry product discovery and development programs. Unique among biotechnology companies is LSB's integration of technologies to discover, analyze, manufacture and find new applications for proteins - the functional units of all biological processes.

Genomics companies have focused on deciphering genetic information, providing an initial but only partial understanding of biological processes. LSB's proprietary protein technologies can enable the transformation of genomic information into products such as drug targets, therapeutics, diagnostics for drug efficacy and toxicity, and traits for agricultural crops. Large Scale Biology has gone beyond the "genomics" realm in its business model and developed ways to integrate the discovery of gene function with quantitative protein analysis and protein manufacturing. This integration of technology platforms favorably positions LSB as a leading provider of valuable content to industry leaders in the fields of diagnostics, therapeutics, vaccines and agribusiness.

LSB was founded in 1987 with the goal of commercializing its proprietary GENEWARE viral vector system - a novel technology for gene expression. Using safe RNA viruses to transiently express genes in non-recombinant plants, LSB has positioned itself in the industry to provide cost-effective manufacturing and purification of diverse protein and peptide products. The same technology can be applied to the expression of libraries of foreign genes in an automated, high-throughput format to discover the function of genes with unparalleled efficiency. The GENEWARE system and associated proprietary technologies form the basis for LSB's functional genomics, biomanufacturing and a variety of proprietary products under development.

From its foundation, LSB understood the need to integrate functional genomic and protein manufacturing expertise with quantitative protein analysis and informatics to become a world-leader in the protein field. In 1999, LSB acquired a privately held pharmaceutical proteomics company originally founded in 1985. Large Scale Proteomics Corporation (a wholly

owned subsidiary of Large Scale Biology Corporation) is an industry leader in identifying and characterizing proteins in all types of biological samples for the discovery and development of new and more effective therapies, diagnostics, and agricultural products.

"Proteomics" is the study of the entire complement of proteins expressed in a cell, tissue, or organism. Proteomics can significantly improve drug discovery and development because most illness is associated with imbalances among, or malfunctions of, proteins. Only a small fraction of diseases can be attributed to the presence of a defective gene. Unlike classical genomics approaches that discover genes that may relate to a disease, LSP has developed a proprietary system called the ProGEx module for directly characterizing proteins associated with disease. Using this same technology, LSP can characterize the effects of candidate drugs intended to reverse a disease process, and to determine the degree to which this objective is achieved free of adverse side effects.

LSB and LSP have protected their many discoveries through an extensive portfolio of domestic and foreign patents and have developed commercial alliances and partnerships to exploit the value of their technologies. LSB and LSP scientists and engineers focus on the development and application of resources to help clients meet their objectives as well as the development of our own proprietary products for subsequent partnering with industry leaders.

A combined staff of 140 professionals operates from three locations in the United States, with a network of collaborators and affiliates throughout the US and Europe. Company headquarters, R&D laboratories and its Genomics division are located in Vacaville, California about 60 miles northeast of San Francisco. Process development and biomanufacturing take place in Owensboro, Kentucky, and LSB's Large Scale Proteomics Corporation subsidiary is located in Germantown, Maryland.

In August, 2000, LSB completed an initial public offering (IPO) of 5 million shares of common stock and now trades on the NASDAQ under the symbol LSBC.

### **Leadership - Large Scale Biology Corporation**

*Robert L. Erwin*, Chairman of the Board and Chief Executive Officer, founded LSB™ and has served as a director and officer since 1987. Mr. Erwin is the former chairman of the State of California Breast Cancer Research Council and currently serves on the University of California President's Engineering Advisory Council. He is Chairman of the Supervisory Board of Icon Genetics AG. As a co-founder of Sungene Technologies Corp., Mr. Erwin served as Vice President of Research and Product Development from 1981 through 1986. He has served on the Biotechnology Industry Advisory Board for Iowa State University. Mr. Erwin received his M.S. degree in Genetics from Louisiana State University and is an inventor on several LSB patents.

*David R. McGee, Ph.D.*, a co-founder of LSB and Senior Vice President and Chief Operating Officer, has been an officer since 1987. Prior to joining LSB, Dr. McGee was Vice President of Operations at Sungene Technologies Corporation from 1983 to 1987. Dr. McGee received his Ph.D. in Genetics from Louisiana State University and served as a faculty instructor of zoology and genetics at Louisiana State University.

*Laurence K. Grill, Ph.D.*, a co-founder of LSB and Senior Vice President, Research and Development, has served as an officer since 1987. Dr. Grill was the Manager of Plant Molecular Biology for Sandoz Crop Protection Corp. from 1984 to 1987 and Senior Research

Scientist in the Department of Molecular Biology at Zoecon Research Institute from 1980 to 1984. He received his Ph.D. from the University of California at Riverside with an emphasis on the molecular basis for viral gene expression in plants.

*R. Barry Holtz, Ph. D.*, Senior Vice President, Biopharmaceutical Manufacturing, has served the company as an officer since 1989 upon the acquisition of Holtz Bio-Engineering, which was founded in 1980. Dr. Holtz was a co-founder and Director of Research for MFI, Inc., the largest manufacturer of microencapsulated nutrients for agriculture and Director of Fundamental Research at Foremost-McKesson, Inc. Dr. Holtz received his Ph.D. in Biochemistry from Pennsylvania State University and served as Assistant Professor in the Department of Food Science and Nutrition at Ohio State University.

*Daniel Tusé, Ph.D.*, has been an officer of LSB since he joined the Company in 1995 as Vice President, Pharmaceutical Development. Dr. Tusé manages the company's pharmaceutical design and development programs, including LSB's novel vaccines and immunotherapeutics initiatives. Prior to joining LSB, Dr. Tusé was Assistant Director of SRI International's (Menlo Park, Calif.) Life Sciences Division. In his 17 years at SRI, Dr. Tusé developed extensive R&D experience in pharmaceuticals and specialty chemicals, serving an international list of clients. Dr. Tusé received his Ph.D. in Microbiology (1980, *cum laude*) with a minor in Toxicology from the University of California, Davis.

*John S. Rakitan*, a co-founder of LSB, Senior Vice President & General Counsel and Secretary, has served as an officer since 1988. Prior to joining LSB, Mr. Rakitan was an attorney in private practice. Mr. Rakitan received his J.D. degree from the University of Notre Dame.

*Michael D. Centron*, Treasurer, has served as Controller since 1988 and was elected as Treasurer in 1991. Mr. Centron was Audit Supervisor for Varian Associates from June 1985 through July 1988, and he also worked for Arthur Young and Co. (currently Ernst & Young). Mr. Centron is a certified public accountant and received his M.B.A. degree from the University of California at Berkeley.

*Guy della-Cioppa, Ph.D.*, is an officer of the company and currently serves as Vice President, Genomics. Prior to joining the company in 1989, Dr. della-Cioppa worked for Monsanto Company in St. Louis, MO from 1984-1989 and was an NIH Postdoctoral Fellow at the Worcester Foundation for Experimental Biology in Shrewsbury, MA from 1983-1984. He received his Ph.D. in Biology from the University of California, Los Angeles.

*William M. Pfann* joined Large Scale Biology in August 2000 as Senior Vice President Finance and Chief Financial Officer. Mr. Pfann was formerly with PricewaterhouseCoopers LLP from 1969 to July 2000, most recently as the Risk Management Partner for the Western Region. He served in a number of management roles at PwC, including leader of the firm's Silicon Valley audit practice, National Director of the networking and communications sector and Managing Partner of the Northern California emerging business group, as well as Partner-in-Charge of the Oakland and Walnut Creek, California offices. Mr. Pfann received a B.S. degree from the University of California, Berkeley, in Business Administration and an MBA in Accounting from Golden Gate University.

**[back to index](#)**

---

## Large Scale Proteomics Corporation

### Leadership - Large Scale Proteomics Corporation

*N. Leigh Anderson, Ph.D.*, Chairman, President and CEO of Large Scale Proteomics Corporation (LSP™). Dr. Anderson obtained his B.A. in Physics with honors from Yale and a Ph.D. in Molecular Biology from Cambridge University (England) working with M. F. Perutz as a Churchill Fellow at the MRC Laboratory of Molecular Biology. Subsequently he co-founded the Molecular Anatomy Program at the Argonne National Laboratory (Chicago) where his work in the development of 2-dimensional electrophoresis (2-DE) and molecular database technology earned him, among other distinctions, the American Association for Clinical Chemistry's Young Investigator Award for 1982 and the 1983 Pittsburgh Analytical Chemistry Award. In 1985 Dr. Anderson co-founded LSP (originally Large Scale Biology Corp., Germantown, MD) in order to pursue commercial development and large-scale applications of 2-D electrophoretic protein mapping technology.

*Norman G. Anderson, Ph.D.*, Chief Scientist at LSP. Dr. Anderson has a distinguished record as an inventor. His career includes senior positions at Oak Ridge and Argonne National Laboratories (ORNL and ANL), more than 300 scientific publications, and the receipt of more than 20 prestigious awards in recognition of his work in science and technology. For his invention of the zonal ultracentrifuge, he received the John Scott Medal Award, and for the centrifugal fast analyzer, the Preis Biochemische Analytik für Klinische Chemie from Die Deutsche Gesellschaft für Klinische Chemie for the most outstanding analytical development in clinical chemistry worldwide during a 2-year period. In 1984 ANL awarded him its career patent leader award for the largest number of patents issued to an employee. At that time the commercial value of his inventions in terms of U.S. sales and royalties from foreign licensing were \$250 million and \$1 million, respectively. Dr. Anderson received his degrees at Duke University: a B.A. in Zoology, M.A. in Physiology, and Ph.D. in Cell Physiology. He holds 28 patents.

*Constance Seniff*, Vice President, Operations. Ms. Seniff has managed LSP's operations since 1993. Her background includes thirteen years in international business prior to joining LSP, five abroad in the employ of foreign firms. Ms. Seniff is responsible for helping formulate and implement business development and database commercialization strategies for LSP in coordination with the management of LSP's parent company, Large Scale Biology Corporation. Ms. Seniff has a B.Sc. degree in Business (with honors) from Florida State University.

*Robert J. Walden*, Vice President, Finance at LSP. Mr. Walden joined LSP in 1997 and has served as a director since 1999. He previously served as Vice President of Finance and Administration at Osiris Therapeutics, Inc., and as Chief Financial Officer at the American Type Culture Collection (ATCC). Mr. Walden received his degree in Finance from the University of Maryland.

*Jean-Paul Hofmann, Ph.D.*, Vice President, Software Development at LSP. Dr. Hofmann is a plant geneticist by training, having earned a B.S. in Biology, M.S. in Biochemistry and Genetics, and Ph.D. in Plant Genetics from the University of Orsay, Paris. He has extensive

experience in using 2-DE in agronomic research and in designing analytical software for 1- and 2-D applications. He has held senior scientific positions in industry and research institutes, in the U.S., France and the Ivory Coast.

*John Taylor, Ph.D.*, Vice President, Software Development and Bioinformatics. Dr. Taylor is the principal developer of Kepler™, LSP's analytical software for automated 2-DE pattern analysis. Prior to joining LSB, Dr. Taylor served as computer scientist in the Molecular Anatomy Program at Argonne, and on the research staffs of the University of Chicago and the Armed Forces Institute of Pathology in Washington, D.C. Dr. Taylor received a B.S. in Physics from the University of South Carolina, and a Ph.D. in Nuclear Physics from Duke University.

*Sandra Steiner, Ph.D.*, currently serves as Vice President Proteomics Applications. Prior to joining the Company, Dr. Steiner founded and directed the Molecular Toxicology Group at Novartis in Basel, Switzerland and was a member in several multi-disciplinary drug development project teams. Dr. Steiner received her Ph.D. in Toxicology/Pharmacology from the University of Basel, Switzerland.

**[back to index](#)**

© 2000 Large Scale Biology Corporation. All Rights Reserved Worldwide.